Physics

for the IB DIPLOMA

OPTIONS

John Allum
Christopher Talbot

FREE ONLINE RESOURCES

HODDER EDUCATION

Option A

13 Relativity

ESSENTIAL IDEAS

- Einstein's study of electromagnetism revealed inconsistencies between the theories of Maxwell and Newton's mechanics; he recognised that these theories could not both be correct. Einstein chose to trust Maxwell's theory of electromagnetism but, if Maxwell was right, Einstein saw that this forced extraordinary changes to long-cherished ideas about space and time in mechanics.
- Observers in relative uniform motion disagree on the numerical values of space and time coordinates for events, but agree with the numerical value of the speed of light in a vacuum. Lorentz transformation equations relate the values in one reference frame to those in another. These equations replace the Galilean transformation equations that fail for speeds close to that of light.
- Spacetime diagrams are a very clear and illustrative way of showing graphically how different observers in relative motion to each other have measurements that differ from each other.
- Energy must be conserved under all circumstances, and so must momentum. The relativity of space and time requires new definitions for energy and momentum in order to preserve the conserved nature of these laws under relativistic transformations.
- General relativity is a framework of ideas, applied to bring together fundamental concepts of mass, space and time in order to describe the fate of the universe.

13.1 (A1: Core) The beginning of relativity *– Einstein's study of electromagnetism revealed inconsistencies between the theories of Maxwell and Newton's mechanics; he recognised that these theories could not both be correct. Einstein chose to trust Maxwell's theory of electromagnetism but, if Maxwell was right, Einstein saw that this forced extraordinary changes to long-cherished ideas about space and time in mechanics*

In order to understand the remarkable steps that Einstein took at the start of the twentieth century it is important to appreciate the scientific context in which these steps were taken. Newtonian mechanics had reigned supreme because of the success of Newton's theory in accurately describing motion in the universe. To appreciate how Einstein overturned the Newtonian paradigm, we need to be clear about some of the principles in Newton's model of the universe.

Reference frames

You probably solved many problems in Chapter 2 in which you considered a stationary reference frame and measured or calculated the displacement, velocity and acceleration of different objects relative to a stationary point.

A reference frame is simply a coordinate system that allows a specific value of time and position to be assigned to an event.

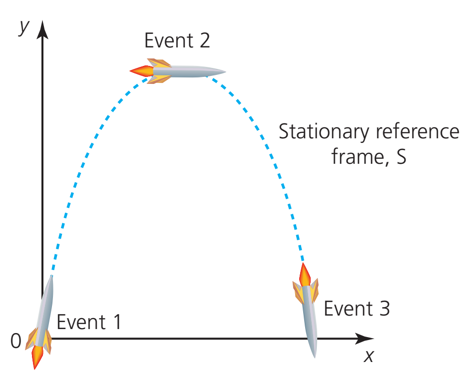


Figure 13.1 Graphical representation of the Earth's reference frame for a rocket in flight

An **event** is an instantaneous incident that occurs at a specific point in space.

Examples of events are a flash of light, the moment when two objects collide and the high point of an object in parabolic flight.

Reference frames are often represented by a set of axes, usually given the label 'S', as shown in Figure 13.1.

To define a reference frame we must specify the origin, the directions of the x and y axes, and the event from which time is started. In the example in Figure 13.1 the obvious reference frame is that of the Earth. However, we could also consider the rocket's reference frame, in which the rocket is stationary and it is the Earth that is seen to move.

The success of Newtonian mechanics is that it allows the accurate calculation of properties such as displacement, velocity, acceleration and time using the equations of motion.

Worked example

1 For the three events shown in Figure 13.1, calculate the *x*, *y* and *t* coordinates of a rocket in freefall with an initial vertical velocity of $400\,\text{m}\,\text{s}^{-1}$ and a horizontal velocity of $100\,\text{m}\,\text{s}^{-1}$.

Event 1: This is the event that defines the origin and also the zero point of time – so $x = 0\,\text{m}$, $y = 0\,\text{m}$ and $t = 0\,\text{s}$, i.e. the coordinates are (0.0 m, 0.0 m, 0.0 s).

Event 2: This is the event defined by the rocket reaching its maximum height. We use the equations of motion to first work out *x*, *y* and *t* as follows:

$$v^2 = u^2 + 2as$$

because $v = 0$ at maximum height:

$$s = \frac{-u^2}{2a}$$

$$s = \frac{-(400^2)}{2 \times (-9.81)}$$

$$s = 8.2 \times 10^3\,\text{m}$$

This gives us the height, so $y = 8.2 \times 10^3\,\text{m}$. For simplicity we will carry forward this value, although it will give a small rounding error. Next calculate the time to reach this point:

$$s = \frac{u + v}{2}t$$

$$t = \frac{2s}{u}$$

$$= \frac{2 \times 8.2 \times 10^3}{400}$$

$$= 41\,\text{s}$$

Since there is no horizontal acceleration it is straightforward to calculate the horizontal position, *x*:

$$s = ut$$

$$= 100 \times 41$$

$$= 4100\,\text{m}$$

Hence the (*x*, *y*, *t*) coordinates of Event 2 are (4100 m, 8200 m, 41 s).

Event 3: This event occurs when the centre of mass of the rocket is the same height as it was originally. The symmetry of parabolic motion means that it occurs at (8200 m, 0 m, 82 s).

1 A car is caught by a speed camera travelling at $35.0\,\text{m}\,\text{s}^{-1}$. If the speed camera photograph is taken at point (0.00 m, 0.0 s) what are the coordinates of the car 23.0 s later?

2 A naughty child throws a tomato out of a car at a stationary pedestrian the car has just passed. The car is travelling at $16\,\text{m}\,\text{s}^{-1}$ and the child throws the tomato towards the pedestrian so that it leaves the car with a speed of $4\,\text{m}\,\text{s}^{-1}$. Explain why the tomato will not hit the pedestrian.

Different reference frames

An **observer** is a theoretical individual who takes measurements from only one specific reference frame. An observer is always stationary relative to their own reference frame and so is sometimes called a **rest observer**.

Have you ever walked up a train as it moves along and wondered what your speed is? If you happen to bang your head twice on bags that stick out too far from the luggage rack as you walk, what are the coordinates of these two head-banging events? The answer to this depends on the reference frame that an observer is taking the measurements from. Here there are three obvious reference frames given by three different observers:

1 An observer taking measurements sitting on the platform as the train moves past.
2 An observer taking measurements sitting on a seat in the train.
3 An observer walking up the train at the same velocity as you.

According to Newton each of these three observers will disagree as to how fast you are moving and also disagree as to your position when you bang your head. However, they will all agree on the time between the two events occurring and the distance you have moved up the carriage between the two events.

Newton's postulates concerning time and space

Newton's description of the universe had many assumptions. Critical for understanding relativity is realising that according to Newton two things must be true:

1 The **universality of time**: all observers agree on the time interval between two events. In particular they must all agree on whether two events are simultaneous or not.
2 The **universality of distance**: all observers agree on the distance between two simultaneous events.

Simultaneous means that the events occur at the same time, i.e. the time interval between the two events is zero.

What do these postulates mean? To understand this better, imagine a universe with a tiny clock placed in the centre of every cubic metre, as shown in Figure 13.2. The first postulate implies that every clock would always be reading the same time and ticking at exactly the same rate. Any observer moving through the universe carrying a clock would find that their clock also reads the same time as the background clocks and must tick at the same rate. If an observer also carries a metre rule with them as they moved around, they would find that it always exactly matched the distance between two adjacent clocks.

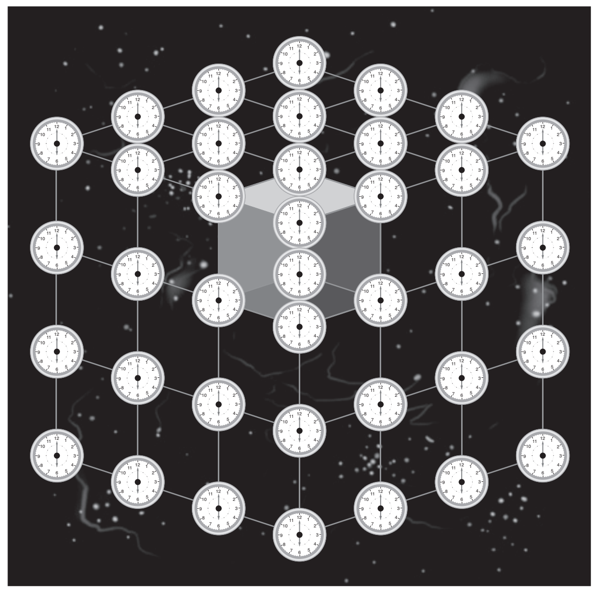


Figure 13.2 A cubic matrix of clocks spreading out regularly throughout space and all reading exactly the same time

Galilean transformations

Whenever we transfer from one reference frame or coordinate system to another, we need to do what is called a **transformation** by applying standard equations. This becomes important when we delve deeper into relativity, so it is worth ensuring that you understand it for Newton's simpler version of the universe.

Galilean transformation equations relate an object's displacement, x, at time t and with velocity u as measured by one observer, to those measured by a second observer travelling with a constant velocity, v, relative to the first observer. The second observer will measure the object's displacement as x' and velocity u' at time t:

$$x' = x - vt$$

$$u' = u - v$$

These equations are given in the *Physics data booklet*.

We use the Galilean transformation equations to allow us to transfer our measurements from one reference frame to another within Newton's model of the universe.

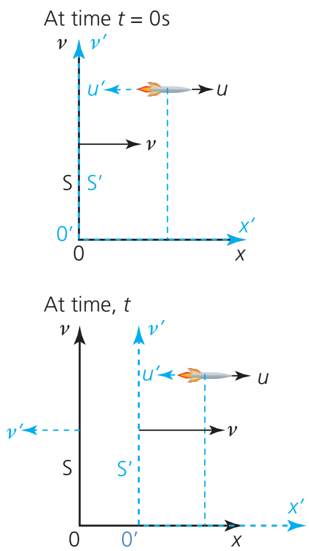


Figure 13.3 Reference frames S and S′ at time $t = 0$ and time t; according to an observer in S the rocket is moving forwards, while an observer in S′ (the dashed blue reference frame) sees it going backwards

Let us look at two reference frames, S and S′, that are moving at constant velocity, v, relative to each other. At time $t = 0$ s, the origins of the two reference frames coincide, so observers in each reference frame agree on the position and time of an event at (0 m, 0 m, 0 s). Figure 13.3 shows this diagrammatically in the top part, by showing the two reference frames on top of each other.

In the bottom part of Figure 13.3 the position of the rocket in space is the same but the positions of the axes, and therefore the reference frames, are different. In reference frame S (solid black lines) the rocket is seen to move forwards while the observer S′ is seen to move forwards even faster; in reference frame S′ (dashed blue lines) the rocket is seen to move backwards while the observer S moves backwards even faster. If you find this confusing try drawing the two reference frames separately so that they are not on top of each other.

Worked example

2 In deep space, rocket A leaves a space station with a constant velocity of $300\,\mathrm{m\,s^{-1}}$. At the same time rocket B travels in the same direction with a constant velocity of $200\,\mathrm{m\,s^{-1}}$.

a What is the distance between rocket A and the space station after one hour?

b According to an observer in rocket B, what is the distance to rocket A after one hour?

c In rocket B's reference frame, how fast would an observer measure the speed of rocket A?

a $x = ut$, where $t = 1 \times 60 \times 60 = 3600\,\mathrm{s}$

$= 300 \times 3600$

$= 1.08 \times 10^6\,\mathrm{m}$

b $x' = x - vt$

$= 1.08 \times 10^6\,\mathrm{m} - (200 \times 3600)$

$= 3.6 \times 10^5\,\mathrm{m}$

c $u' = u - v$

$= 300 - 200$

$= 100\,\mathrm{m\,s^{-1}}$

Assume that the Newtonian model of the universe is correct and use Galilean transformations to answer the following questions. (Note that the answers to some of these questions will contradict the rules of relativity that are introduced in the next section.)

3 In Worked example 2 the rockets travel in the same direction. Use the Galilean transformation equations to calculate the answers to (b) and (c) if the rockets fly off in opposite directions.

4 A rocket travelling at one-tenth of the speed of light away from Earth shines a laser beam forwards into space.

a How fast does an observer inside the rocket measure the light beam photons to be travelling?

b How fast does an observer floating stationary, relative to the Earth, measure the light beam photons to be travelling?

5 Two rockets travelling towards each other are measured by an observer on Earth to each be moving with a speed of 0.6c. How fast does an observer in one rocket think that the other rocket is travelling?

6 If you were in an incredibly fast spaceship that was travelling past a space station at 0.35c and you accelerated a proton inside the ship so that it was travelling forwards through the ship at 0.95c, what speed would an observer in the space station measure the proton to be travelling?

Maxwell's equations

James Clerk Maxwell was a Scottish physicist who is regarded by many as ranking with Newton and Einstein as one of the three greatest physicists of all time. His remarkable work is less well known though, possibly due to its mathematical complexity. Maxwell's outstanding achievement was to link the concepts of electricity, magnetism and optics.

Maxwell brought together four equations that describe how the electrical fields and magnetic fields in a region of space depend on the density of the electrical charge and a property called the current density. They do this by specifying how the fields spread out and how much they curl around. You are not required to have an in-depth understanding of Maxwell's equations or to describe them, but it might be useful to briefly state what each one tells us.

- **Equation 1**: Maxwell's first equation describes how the electrical field varies around a single charge. It describes the repulsion and attraction between like and unlike charges and also states that electrical field lines start on positive charges and finish on negative charges. This is because electric charges are monopoles, meaning that they are either positive or negative.
- **Equation 2**: States that magnetic field lines tend to wrap around things to form closed loops. This is because magnetic poles are always found in north–south pairs and so unlike electric charges are never monopoles.
- **Equation 3**: States that a changing magnetic field will give rise to an encircling electrical field. This is a big step forward because it means that there are two ways that electrical fields are produced – an electrical charge or a changing magnetic field.
- **Equation 4**: States that a changing electrical field will give rise to an encircling magnetic field. This means that a magnetic field can arise not only from a current but also from a changing electrical field.

Maxwell's equations therefore describe a situation in which a changing magnetic field induces an encircling electrical field. The electrical field is also changing so this induces an encircling magnetic field. This new magnetic field is also changing and so the induction goes backward and forward between the electrical and magnetic fields as an oscillation. How strongly they do this is linked to two properties of matter called the **electrical permittivity**, ε, and the **magnetic permeability**, μ, which describe how much charge or current is required to produce a given electrical or magnetic field.

Maxwell and the constancy of the speed of light paradox

Using Maxwell's equations it is possible to derive the speed with which an electromagnetic oscillation propagates in a vacuum using the constants ε_0, the permittivity of free space, and μ_0, the permeability of free space. Maxwell's equations give the value for the speed of propagation of the electromagnetic oscillation in a vacuum as:

$$v = \frac{1}{\sqrt{\varepsilon_0 \mu_0}}$$

$$v = \frac{1}{\sqrt{8.85 \times 10^{-12} \times 4\pi \times 10^{-7}}}$$

$$v = 3.00 \times 10^{8}\,\mathrm{m\,s^{-1}}$$

As you can see, this is the same as the experimental value for the speed of light in a vacuum. The problem is that Maxwell's equations do not give a value that depends on the speed of either the source of the electromagnetic wave or on the speed of the observer – the speed of light, under Maxwell's equations, has to be a constant fixed speed. This is counterintuitive with our experience of moving objects so, before Einstein's work, the solution had usually been to assume that electromagnetic waves travelled through a medium called the **ether** (or aether) and that something needed to be added to Maxwell's equations to account for the speed of the source and the observer.

This would mean that the ether was, by definition, a stationary aspect of the universe from which all observers could calculate their speed by measuring the speed of light. It is comparable to the matrix of clocks suggested in Figure 13.2.

In contrast, Einstein decided that the elegant simplicity of Maxwell's equations was fundamentally correct and that the speed of light in a vacuum was invariant (cannot change). In other words, all observers must always measure the speed of light in a vacuum to be the constant value, *c*, regardless of the observer's own motion.

Nature of Science

Paradigm shift

Einstein's solution was a fundamental shift in our way of thinking about the universe. It was triggered by the apparently simple decision to doubt what appeared to be common sense and to believe in the absolute predictions made by the mathematical model. This decision then led to the theory of relativity.

There was other evidence that not everything was right in the Newtonian model. One example was an attempt to measure the speed of the Earth through the ether using an experiment that is now known as the Michelson–Morley experiment. Its null result – that the Earth was not moving through the ether – caused considerable consternation among both theoretical and experimental physicists.

There have been many paradigm shifts in the development of modern physics: Newton's laws, the use of statistical theory in thermodynamics, the idea that the Earth orbits the Sun. The development of our understanding of atoms has gone through many such shifts – can you describe at least four of them?

Understanding the forces on a charge or a current

Hendrik Lorentz was a Dutch Nobel laureate who was instrumental in providing the framework that Einstein built on to produce special relativity. In particular, Lorentz studied how electrical and magnetic fields would be perceived in different reference frames. Let us look at several different situations and consider the forces that occur and how they are seen to be generated. You will need a sound understanding of the sections in Chapter 5 that explain the forces on charges moving across magnetic fields and to realize that magnetic forces and electrical forces do not act on each other.

A positive charge moving in a static uniform magnetic field

Let us imagine two observers: one in the laboratory's frame of reference and one in the electron's frame of reference. Each observer can measure both magnetic field strength and electrical field strength.

First we will analyze the physics according to the laboratory observer. Before the electron arrives, the observer in the laboratory reference frame measures the static magnetic field produced in the laboratory (but no electrical field because the laboratory is electrically neutral, having an equal number of positive and negative charges). When the electron passes into the magnetic field, this observer will detect a changing electrical field caused by the electron, and a changing magnetic field caused by the changing electrical field (as described by Maxwell's equations). As the electrical field caused by the electron is not interacting with any other electrical field, the laboratory-based observer must assume that it is not involved in the electron's deflection. They must deduce that the force on the electron must be solely due to the interaction between the two magnetic fields.

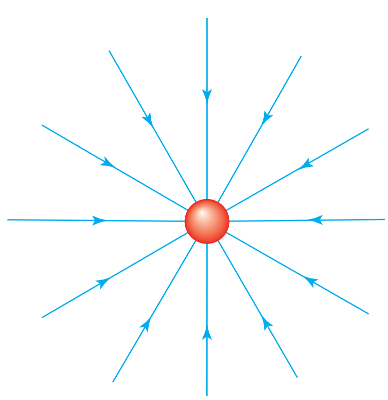
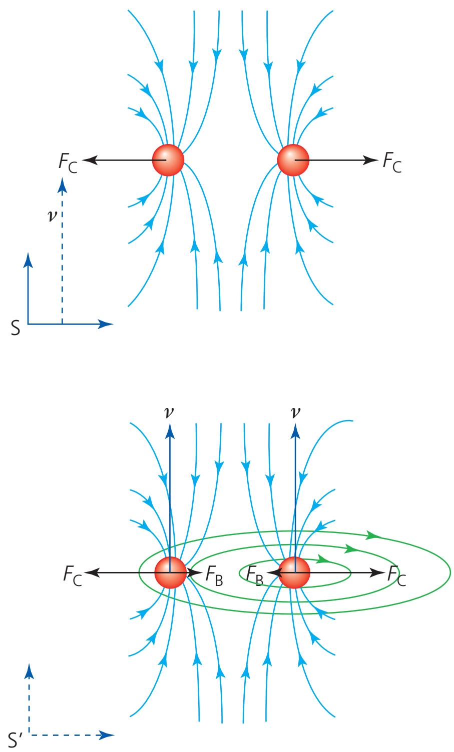


Figure 13.4 An electron, measured by an observer who is stationary relative to the electron, will only have an electrical field around it; there will be no magnetic field

However, in the electron's reference frame the rest observer is stationary with respect to the electron. This means that before the laboratory and the magnetic field arrive, this observer will only detect an electrical field and will be unable to detect a magnetic field, as shown in Figure 13.4. When the laboratory arrives (remember we are now in the electron's reference frame) the electron experiences just the same force but, according to this observer, the force must be purely electrical and is due to an electrical field being produced by the moving magnetic field. Similarly, the observer's magnetic probe



■ **Figure 13.5** A comparison of two reference frames for two electrons moving in parallel. In reference frame S the electrons only experience the repulsive electrical force, F_C. While according to an observer in reference frame S′, the electrons experience both a repulsive electrical force, F_C, and a weaker attractive magnetic force, F_B

also detects the arrival of the magnetic field but the observer will be forced to assume that, because there is only one magnetic field, the force on the electron is not a magnetic interaction.

In all other reference frames a combination of electrical and magnetic fields will combine to produce the force on the electron. In each case the size of the total force is the same, showing us that the electrical and magnetic forces are really just different aspects of the same thing.

Two electrons moving with identical parallel velocities relative to a laboratory

In the electrons' reference frame an observer would measure both electrons to be stationary, so the only interaction that occurs between them is due to the electrical force F_C. This is purely repulsive and can be calculated from Coulomb's law. This is shown in the top diagram in Figure 13.5.

However, in the laboratory the electrons are both moving and so they each generate a magnetic field. If we look at the forces acting on one of the electrons due to the magnetic field around the other electron, we see that the magnetic field attracts the electrons together.

This means that an observer in the laboratory will record the electrical force to be just as strong as before but will also record an attractive magnetic force between the two electrons, so the total force will now be smaller than it was in the electrons' frame of reference.

This means that the total force experienced by the electrons depends on the relative velocity of the frame of reference that is being measured. Lorentz calculated the transformation that makes it possible to easily calculate how this force varies from one reference frame to another using the Lorentz factor, γ. Based on Lorentz's work, Einstein published his 1905 paper on what we now know as 'special relativity'.

ToK Link

When scientists claim that a new direction in thinking requires a paradigm shift in how we observe the universe, how do we ensure their claims are valid?

When any new scientific theory is proposed it is normally presented to the scientific community by publication in a peer-reviewed journal, which provides the first vetting for any new scientific claim. Once a claim is published, other scientists can gain renown in a number of ways: discrediting the new claim through contradictory experimental results or by undermining the theory, or supporting the claim with corroborating experimental evidence and by developing the theoretical argument and the predictions made by the new claim.

To require a 'paradigm shift' means there is something seriously incomplete or wrong with the current theory. So, there should be significant experimental or observational data that is unexplained by the former theory. Of course, for general relativity, the data in support of the theory was minimal at first but it was convincing in the clarity of its predictions.

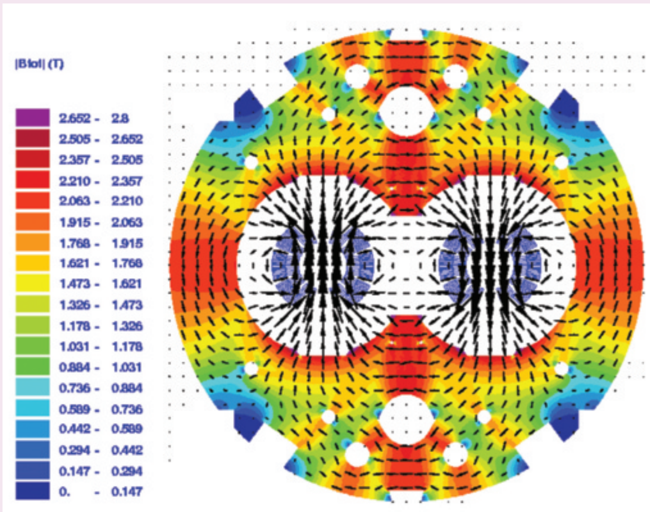
7 What are the values and units of the constants ε_0, the permittivity of free space, and μ_0, the permeability of free space?

8 Explain why two long parallel copper wires, held 1 m apart in a vacuum and each carrying a current of exactly 1 A, will experience an attractive force of $4\pi \times 10^{-7}$ N on each metre of the wire in both the reference frame of the conducting electrons and in the reference frame of the lattice ions.

Utilizations

CERN's Large Hadron Collider

CERN (Conseil Européen pour la Recherche Nucléaire, or in English the European Organization for Nuclear Research) is the largest international collaboration of physicists. It is famous as the location where the Higgs Boson was found in July 2012, but is also celebrated as the birthplace of the worldwide web.



■ **Figure 13.6** Magnetic field pattern of one of CERN's main dipoles in use

CERN's LHC particle accelerator accelerates packets of protons in opposite directions around the 27-km circumference main ring. The protons are travelling very close to the speed of light and so have enormous kinetic energies and momenta.

There are many different types of magnet around the ring, but 60% of the total magnetic energy is used by the superconducting main dipoles – sections of tubing 14.3 m long, that use a current of 11 850 A to produce a vertical 8.33 T field to bend the proton beam into a circle. Test magnets at CERN reached field strengths of 13.5 T in November 2013.

1 The field across each of the ultra-high vacuum beam tubes is shown in Figure 13.6. Describe the forces acting on the proton in a proton cluster passing along one of the beam tubes according to:

a an observer in the proton packet's rest reference frame

b the Earth frame of reference

13.2 (A2: Core) Lorentz transformations *– observers in relative uniform motion disagree on the numerical values of space and time coordinates for events, but agree with the numerical value of the speed of light in a vacuum; Lorentz transformation equations relate the values in one reference frame to those in another; these equations replace the Galilean transformation equations that fail for speeds close to that of light*

Einstein was unhappy with the need to create a unique fundamental reference frame in which an observer will measure the speed of light in a vacuum to be c in all directions. This reference frame would in effect define what was meant by absolutely stationary, and all other reference frames could then have absolute velocities that could be measured from it. Guided by Maxwell's equations he did the opposite – instead of trying to tie down an absolute zero speed, he tied down the maximum speed of the universe.

Unfortunately for Einstein he could only initially get relativistic physics to work in special cases, hence the name **special relativity**. The limiting factor was that special relativity only works in inertial reference frames.

An **inertial reference frame** is one that is neither accelerating nor experiencing a gravitational field. In other words, a reference frame in which Newton's first law of motion is true, so that any object that does not experience an unbalanced force travels in a straight line at constant speed.

To test to see if a reference frame is inertial, carry out a thought experiment and imagine yourself travelling in the reference frame inside a closed box with no means of looking outside of the box. If it is inertial you should be weightless – an apple that is held stationary should remain stationary and a ball that is thrown within the box should travel in a straight line at constant speed across the box.

To work out whether a reference frame is inertial or not, you need to check to see if the reference frame is experiencing any unbalanced external forces or is in a graviational field.

9 Which of the following can be thought of as truly **inertial reference frames**, **almost inertial reference frames** (objects measured over a small distance appear to be travelling at constant velocity) or clearly **not inertial reference frames** (unbalanced forces or gravity are clearly present):
- **a** A rocket stationary in empty space so that it is a long way from any gravitational fields.
- **b** A rocket travelling through empty space in a straight line with constant speed.
- **c** A GPS communication satellite in orbit around the Earth.
- **d** A space probe hovering just above the plasma surface of the Sun.
- **e** A proton travelling close to the speed of light through a *straight* section of tubing in the CERN particle accelerator in Geneva.

Nature of Science

Pure science

Pure science is the study of the facts and rules that describe how the universe works. Pure science contrasts with applied science, which looks at how the laws of science can be applied to solve real-world problems and create real-world applications.

Einstein's work on relativity is a remarkable piece of theoretical pure science – it has been shown to correctly describe the universe at high speeds and energies but also at cosmological scales. It predicts the existence of black holes and gravity waves. However, to non-scientists, and to the politicians responsible for deciding research funding, what use is a pure science like relativity? The answer is often that the spin-offs of fundamental pure science research are unlikely to be predictable. Relativity has resulted in discoveries in other areas such as:
- the expanding universe model and the Big Bang
- atomic and nuclear energy
- aspects of quantum physics
- corrections to GPS satellites

Even so, this is unlikely to reassure everyone. Some politicians may be more impressed by US patent 5280864, which describes a machine for reducing the mass of an object using relativity.

The two postulates of special relativity

First postulate: the laws of physics are identical in all inertial reference frames.

The first postulate does not initially appear to be profound. It could, however, be restated as:
- the stationary observers in different inertial reference frames have equal validity, so no reference frame is more special or unique than any other reference frame
- therefore the universe can have no unique stationary reference frame
- no experiment is possible to show an observer's absolute speed through the universe.

Second postulate: the speed of light *in a vacuum* is a constant, $c = 3.00 \times 10^8\,\mathrm{m\,s^{-1}}$, in all inertial reference frames.

The second postulate supports the evidence of Maxwell's equations but does not appear to make sense with our experience from everyday speeds. It implies that if a rocket in deep space passes a space station at one tenth of the speed of light, and fires a laser beam forwards as it does so, then both an observer in the rocket and an observer on the space station must measure the speed of light to be c, even though they are moving relative to each other. For this to be the case space and time must behave in profoundly different ways from how we have learnt to expect from our real-world experiences – space and time are not in fact like we naively think they are.

Implications of the two postulates

The first implication is that time cannot be invariant in a relativistic universe. Our earlier model of the matrix of clocks is wrong (Figure 13.2). Not only do the clocks read different times and tick at different rates but, for any pair of events, different clocks can record different time intervals. In other words, the time interval between two events does not have to be the same for different observers taking measurements from different inertial reference frames.

There are two reasons for this:

- The first is that information about an event can only travel through the universe at the speed of light; when you think that the event happened depends on where you are relative to the event.
- The second is because time itself becomes distorted by movement – in effect, the faster an observer is moving, the slower time passes.

The second implication is that space is also not invariant – the distances between any matrix of points is not the same for all inertial reference frames and depends on relative motion. The perceived symmetry of relativity can make this point difficult to understand, so we will discuss this in more detail later. Let us first discuss how we measure time.

Clock synchronisation

Imagine you and your friends decide to play a prank on your physics teacher just before the lesson starts by telling them that the headteacher needs to see them urgently. To make matters more interesting, you all have a sweepstake to guess how long into the lesson it will take before the teacher, fuming nicely, returns. To make it fair, everyone has a stopwatch to measure the time. Obviously you all have to synchronise stopwatches – how do you do that?

The answer is fairly simple. You need a single event, like a gun going off at the start of a race, so that everyone starts their stopwatch as soon as they hear the gun go off. The problem is that it takes time for the signal to travel; so what do we do if we want to be incredibly accurate, or if we are trying to synchronise watches that are a long way apart? The solution is shown in Figure 13.7.

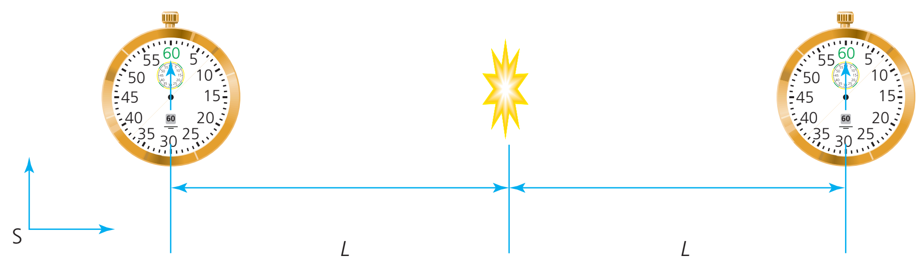


Figure 13.7 Two stopwatches that are some distance apart can be synchronised by a flash of light that is fired exactly in the middle of them, so that the light takes the same amount of time to reach each stopwatch; the stopwatches start as soon as they detect the light signal

For an observer who is stationary with respect to the stopwatches, the two must now read the same time and are therefore synchronised. However, this is not true for an observer who is moving with respect to the stopwatches, along the straight line that joins them, as can be seen in Figure 13.8. This observer, who is moving to the left with speed v relative to the clocks, can prove that, in their reference frame, the light cannot reach the stopwatches simultaneously, that the two in their reference frame physically cannot be synchronised, and that the stopwatch on the left will always read a later time than the one on the right.

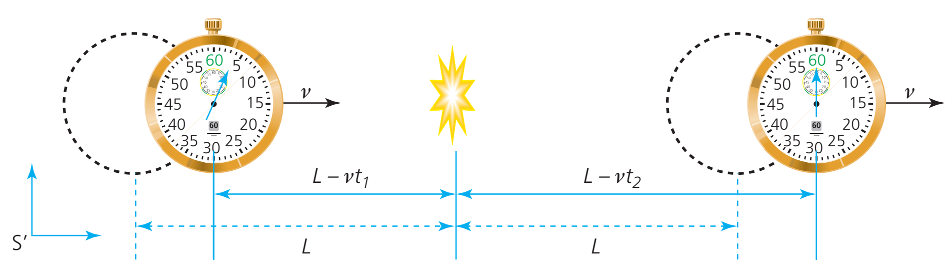


Figure 13.8 An observer who sees the stopwatches moving sideways will see one moving towards the flash and the other moving away from the flash since light is also travelling at *c* according to this observer. The light takes time to travel outwards from the flash and, in this time, one stopwatch is further from the source of the flash than the other and so the clocks cannot be synchronised

Utilizations

Correcting GPS satellites

One of the applications of relativity is the correction of GPS satellites. GPS satellites require the use of atomic clocks to send out very precise time signals that allow the position of the GPS receiver to be measured. The orbit speed of GPS satellites clocks is high enough to result in an error that needs to be corrected for. There is also a correction factor due to the gravitational field and so this application is discussed in much more detail in Section 13.5.

Simultaneous events

How is it possible for a clock to read different times to different observers? This is not just a trick of mathematics but a fundamental aspect of Einstein's relativistic physics. It means that two events that occur some distance apart will appear simultaneous to some observers but not to others.

Imagine a very long train carriage with a mirror at each end. An observer stands exactly in the middle of the carriage and produces a single flash of light (Event 0 in Figure 13.9) that simultaneously sends out light in each direction. These light pulses reflect off the mirrors at the ends of the carriage (Events 1 and 2) and return to the observer in the middle of the carriage (Event 3). A second observer witnesses the experiment from the platform as the train travels past, very fast but at constant velocity.

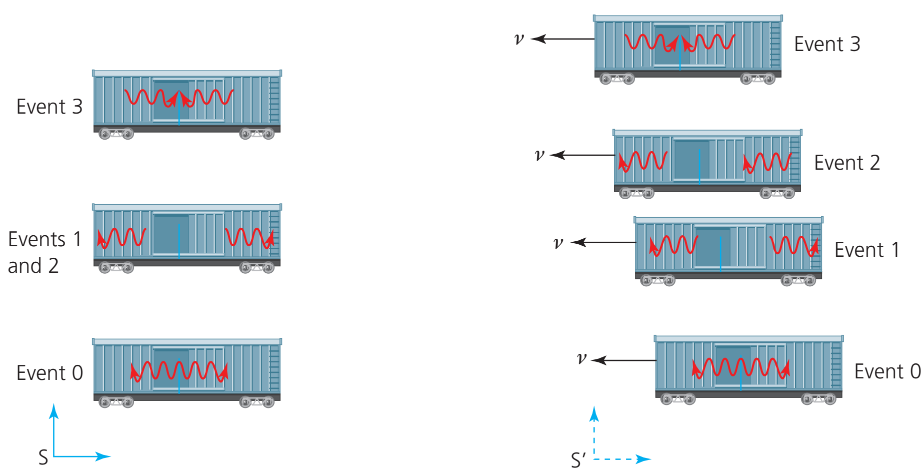


Figure 13.9 Comparison of how two observers in different inertial reference frames record the time of different events. Events 0 and 3 are each a pair of simultaneous events that occur in the same place, so all observers must agree that they are simultaneous. Events 1 and 2 occur in different places so they are fundamentally simultaneous for one observer but fundamentally cannot be simultaneous for the other observer

Observer S sees:

- The pulses are sent out simultaneously (Event 0).
- The pulses reach each the end of the carriage simultaneously (Events 1 and 2).
- The pulses return to observer S simultaneously (Event 3).

Observer S′ sees:

- The pulses are sent out simultaneously (Event 0).
- The pulse that travels down the carriage against the motion of the carriage must arrive at the end of the carriage (Event 1) before the pulse that travels up the carriage (Event 2) because it must have travelled a shorter distance.
- However, S′ still sees the pulses return to observer S simultaneously because the reverse effect is true for the reflected rays (Event 3).

A light clock is sometimes used as a way of comparing observations that are made by observers in two different inertial reference frames. A light clock is a very simple device that reflects a light beam between two parallel mirrors separated by a fixed distance, *d*. The speed of light in a vacuum is constant for all observers but the path length taken by the light varies. See Figure 13.10.

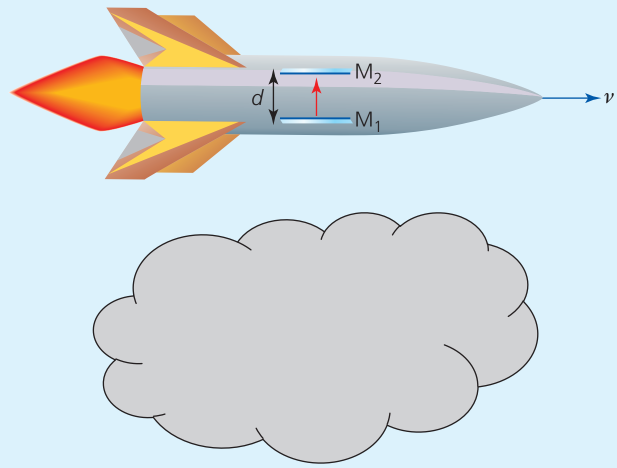


Figure 13.10

10 One of the diagrams in Figure 13.11 shows the path of the light beam as seen by the physicist on the rocket (Rachel), while the other is seen by the physicist hiding in the gas cloud (Gavin), who sees the rocket moving to the right with speed *v*. Which is which?

11 According to Gavin the time that the light pulse takes to travel from M_1 to M_2 is Δt. Therefore, how far does the rocket move sideways in this time?

12 Use a Galilean transformation to work out the speed of the light beam according to Gavin.

13 Using Newtonian physics, how far does the light beam have to travel when reflecting between M_1 and M_2, according to Gavin?

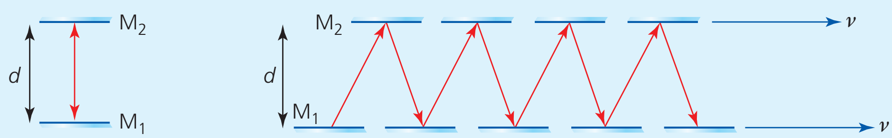


Figure 13.11

14 Gavin sees the rocket moving sideways with speed *v*. In terms of *c* and Δt, how far has the light beam travelled from M_1 to M_2 according to Gavin.

15 According to Rachel in the rocket, the time taken to travel from M_1 to M_2 is $\Delta t'$. Utilizing Pythagoras, use your understanding of the postulates of Newtonian physics to derive an expression for Δt in terms of $\Delta t'$, *v* and *c*.

16 Explain in terms of the constancy of the speed of light why the two observers must disagree about the time it takes for the light beam to travel between M_1 and M_2

All inertial observers must agree that simultaneous events that occur at the same point in space are definitely simultaneous. However, they must disagree about the order of events that occur in different places if they are moving at varying speeds along the line of the two events.

Fundamentally, therefore, the two observers disagree about the order that events must occur in. This means that our understanding of what time is has been incorrect. Time is not something that is intrinsically there in the universe. Instead it is something that is formed by events and something that must vary depending on both an observer's position and motion relative to the events. This is why relativity talks about **spacetime** – the fabric of the universe. We can still take measurements of what we think of as 'time' but what we are actually doing is taking a measurement of spacetime, and this intrinsically means we are taking a measurement that combines elements of both time and space.

So if time is not a fixed, unvarying quantity in a relativistic universe, what quantities can we rely on?

Invariant quantities

Einstein soon realised that some quantities are still invariant; they do not change as a result of relative motion and position. These quantities become very important when calculating the changes that do occur, and we need to be absolutely clear that we can identify these before we can solve harder problems in special relativity. The first invariant quantity we have already stated is the speed of light in a vacuum. We will introduce four more now and further invariant quantities later.

Spacetime interval, Δs^2

In Newton's universe, time and space are both invariant – they have fixed intervals (seconds, metres etc.) that do not vary throughout either space or time. This means that we can measure space and time independently. In relativity, space and time are intrinsically linked into a single concept called spacetime. The interval between two events across spacetime is invariant – different inertial observers can measure different times between events and different distances between events but they must all measure the same interval across spacetime. We call this the spacetime interval, Δs^2. It is given by the formula:

$$(\Delta s)^2 = c^2(\Delta t)^2 - (\Delta x)^2$$

This equation is not given in the *Physics data booklet*.

Note that Δs^2 can be positive, zero or negative. Unfortunately two conventions exist and so it is sometime written as: $(\Delta s)^2 = (\Delta x)^2 - c^2(\Delta t)^2$.

Rest mass, m_0

The rest mass is the mass of an object as measured by an observer who is stationary relative to the object.

Proper time interval, Δt_0

A proper time interval is defined as the time interval between two events as measured by an inertial observer who sees both of the events occur in the same place relative to that observer. The reason for this is that the clock can be placed at the exact position of both events without needing to be moved, so the measurement of spacetime involves no spatial element.

$$(\Delta s)^2 = c^2(\Delta t_0)^2 - 0^2$$

$$(\Delta s)^2 = c^2(\Delta t_0)^2$$

This equation is not given in the *Physics data booklet*.

You need to try to imagine yourself in the reference frame of each observer – are the *x*, *y* and *z* coordinates of the two events the same? If they are then this observer measures the proper time between the two events.

17 In a laboratory, an electron is accelerated by a potential difference of 100 kV. Its speed is then measured by timing how long it takes to pass between two different points measured in the laboratory as being 5.00 m apart. Is the observer in the electron's reference frame or the observer in the laboratory reference frame recording proper time?

18 A rod measured in its rest frame to be one metre in length is accelerated to 0.33*c*. The rod is then timed as it passes a fixed point. Is the observer at the fixed point or the observer travelling with the rod measuring proper time?

19 The same rod is timed by both observers as it travels between two fixed points in a laboratory. If the observers are recording when the front of the rod passes each fixed point, is either observer measuring the proper time?

20 In a third experiment the two observers start timing when the front of the rod passes the first point and stop timing when the end of the rod passes the second point. Is either observer measuring the proper time?

Proper length, L_0

Proper length is defined as the distance measured been two points in space as measured by an inertial observer who is stationary relative to the two points. Similarly, this means that the observer is taking a measurement of just the distance aspect of spacetime because the object being measured is always in that position for that observer; there is no measurement of the time element of spacetime.

$$(\Delta s)^2 = c^2 0^2 - (\Delta x)^2$$

$$(\Delta s)^2 = -(\Delta x)^2$$

This equation is not given in the *Physics data booklet*.

The fact that Δs^2 is negative is to do with how Δs^2 is defined.

Many students struggle to understand this concept initially. In your head imagine joining the two points where the events occur with a piece of string. Now imagine the piece of string from the viewpoint of each observer. If the string is stationary for an observer then they are measuring the proper length. Secondly, an observer will only measure a proper length to be zero if the two events are both simultaneous and occur in the same place.

21 In a laboratory, a proton is accelerated by a potential difference of 100 kV. Its speed is then measured by timing how long it takes to pass between two different points measured in the laboratory as being 5.00 m apart. Is an observer in the electron's reference frame or the observer in the laboratory reference frame recording the proper length between the two points?

22 A rod measured in its rest frame to be one metre in length is accelerated to 0.33*c*. The rod is then timed as it passes a fixed point. Is the observer at the fixed point or the observer travelling with the rod measuring proper length between the start and finish events?

23 The same rod is timed by both observers as it travels between two fixed points in the laboratory. If the observers are recording when the front of the rod passes each fixed point, is either observer measuring the proper length for the distance between the start and finish events?

24 In a third experiment the two observers start timing when the front of the rod passes the first point and stop timing when the end of the rod passes the second point. Is either observer measuring the proper length for the distance between the start and finish events?

Worked examples

3 A single laser pulse is made to trigger two explosion events as it travels through a long vacuum tube. The two events are 99 m apart and the time for the light to travel this distance is 3.3×10^{-7} s. What is the spacetime interval between the two events?

$$(\Delta s)^2 = c^2(\Delta t)^2 - (\Delta x)^2$$

$$= (3.0 \times 10^8)^2 \times (3.3 \times 10^{-7})^2 - 99^2$$

$$= 0.0\,\text{m}^2$$

The spacetime interval for any two events linked by a photon travelling in a vacuum is always zero. Two events linked by an object travelling slower than *c* will have a positive spacetime interval, while two events that are too far apart for a photon to travel between the two events in the time interval between them have a negative spacetime interval.

4 What is the spacetime interval for an electron that is fired with a kinetic energy of 10.0 keV across a gap of 5.0 m?

To make this problem easier it is broken down into three stages. First we use conservation of energy (kinetic energy = electrical potential energy) to calculate the speed of the electron. The charge on the electron, *e*, and the mass of the electron, m_e, are given in the *Physics data booklet*.

$$KE = \frac{1}{2}mv^2 = qV$$

$$v^2 = \frac{2qV}{m}$$

$$v = \sqrt{\frac{2qV}{m}}$$

$$= \sqrt{\frac{2 \times (1.60 \times 10^{-19}) \times (10.0 \times 10^3)}{9.11 \times 10^{-31}}} = 5.93 \times 10^7\,\text{m}\,\text{s}^{-1}$$

Next, it is a simple matter to calculate the time interval:

$$\Delta t = \frac{\Delta x}{v}$$

$$= \frac{5.00}{5.93 \times 10^7} = 8.44 \times 10^{-8}\,\text{s}$$

Finally, we can use the equation given earlier to calculate the spacetime interval. It gives a positive answer (and note that the units are m^2):

$$(\Delta s)^2 = c^2(\Delta t)^2 - (\Delta x)^2$$
$$= (3.00 \times 10^8)^2 \times (8.44 \times 10^{-8})^2 - 5.00^2$$
$$= 6.16 \times 10^2\,\text{m}^2$$

The fact that the spacetime interval between any two events is constant for all observers allows us to calculate how long a time an observer travelling in the electron's reference frame will record between the two events. In this reference frame the electron is stationary and the start and finish lines move towards it with the start and finish events occurring at the electron. This means that this observer is recording *proper time* and $\Delta x' = 0$:

$$(\Delta s)^2 = c^2(\Delta t')^2 - (\Delta x')^2$$
$$(\Delta t')^2 = \frac{(\Delta s)^2}{c^2}$$
$$\Delta t' = \sqrt{\frac{6.2 \times 10^2}{(3.00 \times 10^8)^2}}$$
$$= 8.27 \times 10^{-8}\,\text{s}$$

The electron therefore experiences a slightly slower time for these two events than the observer taking measurements in the laboratory.

An alternative way to do these sorts of calculations is to use transformation equations. The invariance of the spacetime interval means that it can be written as:

$$(\Delta s')^2 = (\Delta s)^2$$
$$c^2(\Delta t')^2 - (\Delta x')^2 = c^2(\Delta t)^2 - (\Delta x)^2$$

These equations are not given in the *Physics data booklet*.

If both observers agree to use the same event as the origin of their reference frames then $\Delta x = x$, and $\Delta t = t$, $\Delta x' = x'$, and $\Delta t' = t'$, so the previous equation can also be written:

$$(ct')^2 - (x')^2 = (ct)^2 - (x)^2$$

This equation is given in the *Physics data booklet*.

The Lorentz factor, γ

Einstein used the mathematics of Lorentz to describe how spacetime must change at different relative motions. First meet the Lorentz factor, γ, which allows us to simplify many of the equations that are used in special relativity:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

This equation is given in the *Physics data booklet*.

If you have a programmable calculator it is worth entering this into the equations because it will save time later and allow you to quickly calculate γ for different values of v. In many of the problems that you will work through, you will find that the relative velocity, v, is given as a proportion of c. This makes the calculations simpler.

The Lorentz factor, γ, is a scaling factor – both a mathematical expression of the theory and a convenient way to calculate the magnitude of the changes that have to be made to nominal space and time values or intervals, when looking at events from different reference frames (i.e. different velocities of an 'observer').

So, the size of the Lorentz factor obviously depends on the amount of difference between the chosen frame's velocity and the speed of light, c, but remember that the absolute fastest velocity possible *is* c so the Lorentz factor becomes asymptotically huge as speeds approach c, a fact that has fascinating consequences.

Worked examples

5 Calculate the value of γ for a particle travelling at:
a 50% of the speed of light in a vacuum
b 99% of the speed of light in a vacuum.

a $\gamma = \dfrac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}}$

$= \dfrac{1}{\sqrt{1 - \dfrac{(0.5c)^2}{c^2}}}$

$= \dfrac{1}{\sqrt{1 - 0.25}}$

$= 1.2$

b $\gamma = \dfrac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}}$

$= \dfrac{1}{\sqrt{1 - \dfrac{(0.99c)^2}{c^2}}}$

$= \dfrac{1}{\sqrt{1 - 0.99^2}}$

$= 7.1$

Hopefully, you can see from these examples that the value of γ must always be bigger than 1, and that it has no units.

6 Calculate the value of v when $\gamma = 1.75$.

It is standard practice to give an answer in terms of c.

$\gamma = \dfrac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}}$

$\gamma^2 = \dfrac{1}{1 - \dfrac{v^2}{c^2}}$

$1 - \dfrac{v^2}{c^2} = \dfrac{1}{\gamma^2}$

$\dfrac{v}{c} = \sqrt{1 - \dfrac{1}{\gamma^2}}$

$= \sqrt{1 - \dfrac{1}{1.75^2}}$

$= 0.821$

$v = 0.821c$

25 What is the value of γ for the relative velocities of:
a $v = 0.1000c$
b $v = 0.75c$
c $v = 0.90c$
d $v = 0.95c$

26 Sketch a graph of γ against v for velocities from 0 to $0.999c$.

27 What is the value of v that has the following values of γ:
a $\gamma = 1.00$
b $\gamma = 1.15$
c $\gamma = 2.00$
d $\gamma = 4.00$

Lorentz transformations

In Newton's model of the universe, we used Galilean transformation equations to move from one reference frame to another, allowing us to change one coordinate into another. In Einstein's relativistic universe we must instead use the Lorentz transformation equations.

$$x' = \gamma(x - vt)$$

$$t' = \gamma\left(t - \frac{vx}{c^2}\right)$$

These two equations are given in the *Physics data booklet* and are used to transform x coordinates and t coordinates where the origins of the two reference frames coincide at $t = 0$ s. Only the x dimension is affected by the transformations, so even more complicated problems can be rotated to simplify them.

Alternatively, the equations are also stated for intervals as a measured length, Δx, and a measured time, Δt, between two events.

$$\Delta x' = \gamma(\Delta x - v\Delta t)$$

$$\Delta t' = \gamma\left(\Delta t - \frac{v\Delta x}{c^2}\right)$$

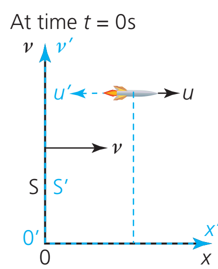
These two equations are given in the *Physics data booklet*.

Worked examples

7 Let's have another look at the example we looked at in Figure 13.3. At time $t = 0$ s the two reference frames, S and S′, coincide, and we will only consider relative movement in the x-spatial dimension. For simplicity only the diagram at time t is considered.

The **light year** (ly) is a unit of distance. It is equal to the distance travelled by light in a vacuum in 1 year. That is:

1 ly = c × 1 y



Let's suppose that, according to a rest observer in reference frame S, the rocket reaches a point 20 light years away after 30 years. This gives (x, t) coordinates for the rocket as (20 ly, 30 y). If we move to a reference frame S′ that is moving at 0.5c relative to S, then what are the coordinates of the rocket according to S′?

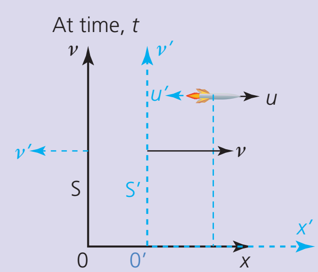
We have already calculated $\gamma = 1.2$ for $v = 0.5c$.

$x' = \gamma(x - vt)$

= 1.2 (20 ly – 0.5c × 30 y)

= 1.2 (20 ly – 15 ly)

= 6.0 ly



$$t' = \gamma\left(t - \frac{vx}{c^2}\right)$$
$$= 1.2\left(30\,y - \frac{0.5c \times 20\,\text{ly}}{c^2}\right)$$
$$= 1.2\,(30\,y - 10\,y)$$
$$= 24\,\text{y}$$

Therefore, according to an observer in reference frame S′, the rocket has only travelled 6 ly in 24 years, which means that it is travelling at only 0.25 *c*. This example is straightforward because the units being used allow *c* to be cancelled easily.

8 The distance between two events as seen by one observer is 250 m with the two events occurring 1.7×10^{-6} s apart, with the event on the left occurring before the event on the right. What is the distance and time interval between the two events as measured by a second observer travelling at 0.75*c* to the right according to the first observer?

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$
$$= \frac{1}{\sqrt{1 - 0.75^2}}$$
$$= 1.51$$
$$x' = \gamma(x - vt)$$
$$= 1.51 \times (250\,\text{m} - 0.75 \times (3.0 \times 10^8\,\text{m s}^{-1}) \times 1.7 \times 10^{-6}\,\text{s})$$
$$= -200\,\text{m}$$
$$t' = \gamma\left(t - \frac{vx}{c^2}\right)$$
$$= 1.51 \times \left(1.7 \times 10^{-6}\,\text{s} - \frac{0.75 \times (3.0 \times 10^8\,\text{m s}^{-1}) \times 250\,\text{m}}{(3.0 \times 10^8\,\text{m s}^{-1})^2}\right)$$
$$= 1.6 \times 10^{-6}\,\text{s}$$

In other words, the order of the events is the same for the second observer but, from this reference frame, their spatial positions are almost reversed.

For each of these problems assume one dimensional motion and assume that, in each case, the observers start timing when the origins of the two reference frames coincide.

28 Imagine a situation where a rocket passes the Earth at 0.5*c*. There are two observers – one in the Earth frame of reference and the other in the rocket frame of reference.

- **a** What is the value of γ?
- **b** A star explosion event occurs at a point 20 light years from the Earth. The rocket passes the Earth heading towards the star. According to the Earth-based observer the rocket passes the Earth 20 years before the light arrives. What are the x' and t' coordinates of the explosion event for the observer in the rocket's reference frame?

29 From Earth, the Milky Way galaxy is measured to be 100 000 light years in diameter, so the time taken for light to travel from one side of the Milky Way to the other is 100 000 years. What is the diameter of the Milky Way for an observer in a distant galaxy moving at a speed of 0.2*c* away from Earth? Assume that they are travelling in the same plane as the measured diameter of the Milky Way.

30 According to Earth-based astronomers a star near the centre of the Milky Way exploded 800 years before a star 2000 ly beyond it. How much later is the second explosion according to a rocket travelling towards the explosions at 0.2*c*?

31 In a laboratory an electron is measured to be travelling at 0.9*c*. According to an observer in the lab at $t = 9.6 \times 10^{-9}$ s it is at a position of $x = 2.6$ m down the length of a vacuum tube. Calculate the value of the Lorentz factor and use it to work out the time and position of the electron according to an observer in the electron's reference frame.

32 Two inertial observers are travelling with a relative velocity of 0.8*c* and both see two events occur. According to one observer the events occur 4.2 m apart and with a time interval of 2.4×10^{-8} s between them. According to the other observer, what are the spatial ($\Delta x'$) and temporal ($\Delta t'$) intervals between the two events?

Demonstrating that events can be simultaneous for one observer but not another

Remember that:

If two events are simultaneous and the two events occur in the same place then they must be simultaneous for all observers.

This is because the spacetime interval is zero for all observers, and so all observers must agree that the two events occur in the same place and, therefore, occur simultaneously.

However,

Two events that occur in different places may be simultaneous to one observer but not to another.

To prove this let us look at the Lorentz transformation equations. Let's suppose that for observer S the two events are simultaneous, $\Delta t = 0$, but occur a length, L, apart, $\Delta x = L$. According to a second inertial observer S′ moving at constant velocity, v, parallel to the length, L, the two events will occur with a time interval of:

$$\Delta t' = \gamma\left(\Delta t - \frac{v\Delta x}{c^2}\right)$$
$$= -\gamma\left(\frac{vL}{c^2}\right)$$

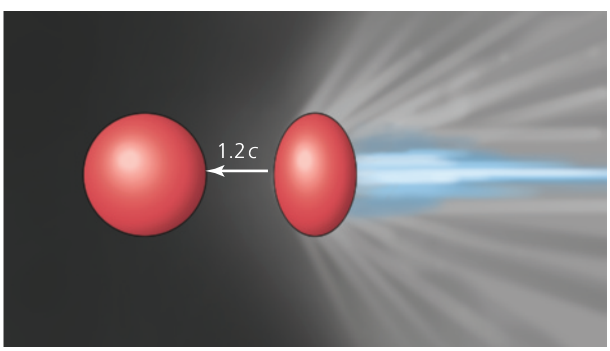
This equation is *not* given in the *Physics data booklet*.

Because γ and c are both non-zero, any observer S′ must record a time interval between the two events for any non-zero values of v and L, and so the two events cannot be simultaneous for the observer.

Velocity addition transformations

One aspect of special relativity that is explained more fully at Higher Level is that it is not possible for any known particles to travel faster than the speed of light in a vacuum. In Newton's universe, when two particles are travelling towards each other at 0.6c, then they would each perceive the other particle to be approaching them at 1.2c (because $u' = u - v$). See Figure 13.12.

Figure 13.12



This is not possible in Einstein's spacetime; instead, we must use a more complicated transformation equation:

$$u' = \frac{u - v}{1 - \dfrac{uv}{c^2}}$$

This equation is given in the *Physics data booklet*. As seen from an external reference source, u is the velocity of object 1, v is the velocity of object 2, and u' is the velocity of object 1 as seen by an observer who is at rest with object 2. Remember that u' must always be less than c.

Worked examples

9 Two particles are seen from an external reference frame to be travelling towards each other, each with a velocity of 0.60c (Figure 13.13). An observer with one particle measures the velocity of the other particle; what do they measure its speed to be?

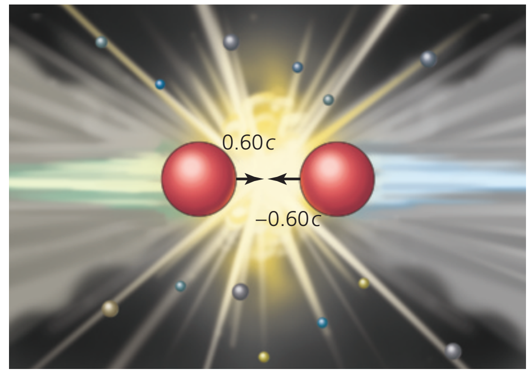
$$u' = \frac{u - v}{1 - \frac{uv}{c^2}}$$

$$= \frac{0.60c - (-0.6c)}{1 - \frac{(0.60c \times (-0.6c))}{c^2}}$$

$$= \frac{1.2c}{1 - \left(-\frac{0.36c^2}{c^2}\right)}$$

$$= \frac{1.2}{1.36}c$$

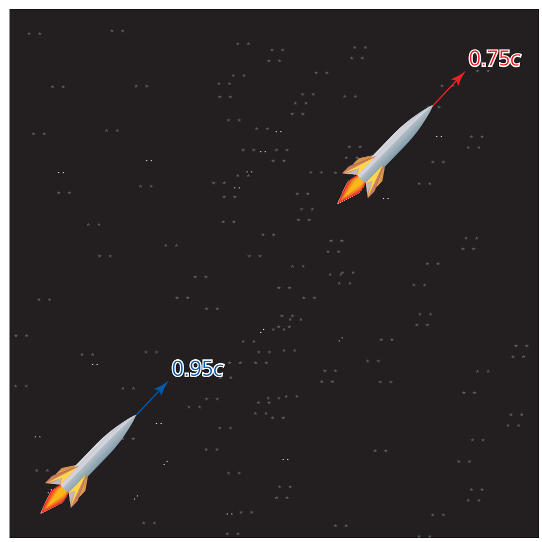
$$= 0.88c$$



■ **Figure 13.13**

It is very easy to miss out the negative signs when doing this calculation. Remember that u and v are both vectors and so can be positive or negative depending on their direction.

10 Two rockets are observed from an external reference frame to be travelling in the same direction – the first is measured to be travelling through empty space at 0.75c, and a second rocket is sent after it is measured to be travelling at 0.95c (Figure 13.14). How fast would an inertial observer travelling with the first rocket measure the approach of the second rocket to be?



■ **Figure 13.14**

$$u' = \frac{u - v}{1 - \frac{uv}{c^2}}$$

$$= \frac{0.95c - 0.75c}{1 - \frac{(0.95c \times 0.75c)}{c^2}}$$

$$= \frac{0.2c}{1 - \left(\frac{0.71c^2}{c^2}\right)}$$

$$= \frac{0.2}{0.29}c$$

$$= 0.70c$$

33 A rocket travelling at one-tenth of the speed of light away from Earth shines a laser beam forwards into space.

a An observer inside the rocket accurately measures the speed of the light beam photons. What value would you expect them to obtain?

b An observer floating stationary, relative to the Earth, also accurately measures the light beam photons. What value will they obtain?

34 Two rockets are flying towards each other; each are measured by an observer on Earth to be moving with a speed of 0.7c. How fast does an observer in one rocket think that the other rocket is travelling?

35 If you were in an incredibly fast spaceship that was travelling past a space station at 0.35c and you accelerated a proton inside the ship so that it was travelling forwards through the ship at 0.95c, relative to the ship, what speed would an observer in the space station measure the proton to be travelling?

36 In an alpha-decay experiment the parent nucleus may be considered to be stationary in the laboratory. When it decays, the alpha particle travels in one direction with a velocity of 0.7c while the daughter nucleus travels in exactly the opposite direction at 0.2c. According to an observer travelling with the daughter nucleus, calculate how fast the alpha particle is travelling.

37 In a beta-decay experiment an electron and an anti-neutrino that are produced happen to travel away in exactly the same direction. In the laboratory reference frame the anti-neutrino has a velocity of $0.95c$ and the electron has a velocity of $0.75c$. What is the anti-neutrino's velocity according to an observer travelling in the electron's reference frame?

38 Protons in CERN's LHC travel in opposite directions around the ring at over $0.9990000c$. According to an observer travelling with one group of protons, how fast are the approaching protons travelling?

39 Two light beams are travelling in exactly opposite directions. According to a laboratory observer their relative velocity is $2c$. How fast would an observer travelling in the reference frame of one of the light beam's photons measure the speed of the approaching light beam's photons to be?

40 In a space race two spaceships pass a mark and are measured by the race officials at the mark to be travelling in the same direction and travelling at $0.6c$ and $0.7c$ respectively. According to the faster spaceship, how fast is the other ship travelling?

Time dilation

It may have seemed surprising in Worked example 10 that observers in the two frames of reference travelling with the rockets perceive that they are apparently moving towards each other so fast ($0.70c$) when Newtonian mechanics suggests that they should only be moving with a relative velocity of $0.20c$. The reason is to do with how time is being slowed by their relatively high velocity.

The shortest possible time between two events is measured by the inertial observer who measures the proper time interval, Δt_0, between those two events. Any other inertial observer will measure a longer time interval, $\Delta t'$, between the events. This slowing, or stretching, of time due to relative motion is called **time dilation**.

$$\Delta t' = \gamma \Delta t_0$$

This equation is given in the *Physics data booklet*.

The derivation is straightforward and depends on the Lorentz transformation for time. Remember that for an observer measuring proper time the two events must be seen to occur at the same spatial coordinates relative to them, and so $\Delta x = 0$.

$$\Delta t' = \gamma\left(\Delta t - \frac{v\Delta x}{c^2}\right)$$

For Δt_0 we know that $\Delta x = 0$, so:

$$\Delta t' = \gamma \Delta t_0$$

Additional Perspectives

The Ives–Stilwell experiment

One of the earliest tests of time dilation was an experiment carried out by Herbert Ives and his assistant G. R. Stilwell in an attempt to disprove special relativity. They used a hydrogen discharge tube that accelerated H_2^+ and H_3^+ ions to high speeds. The ion beam glows when free electrons are absorbed by the ions producing an emission spectrum. They used a concave mirror to produce a reflection of the beam and observed both the original beam and the reflected beam through a spectroscope.

The ions are travelling at high speed so the observed spectrum experiences a classical, or longitudinal, Doppler shift. The effect of this shift is symmetrical on the reflected beam though, so that the blueshift of the original beam is the same size as the redshift of the reflected beam. These two spectral lines are compared with the unshifted original spectral line. Classical theory therefore predicts that three lines with equal spacing are observed (Figure 13.15)

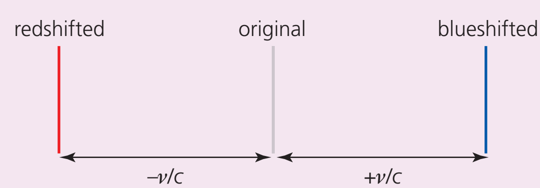
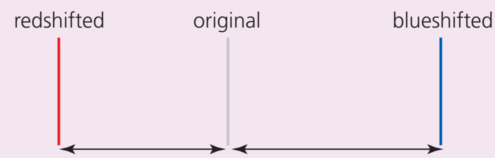


Figure 13.15

However, there is also a much smaller relativistic transverse Doppler-shift effect that was predicted by Einstein in his 1905 paper. This causes the position of the three spectral lines to be additionally shifted by different amounts so that the original line is no longer in the centre (Figure 13.16)



■ **Figure 13.16**

Unfortunately for Ives, who was one of America's most ardent critics of relativity, the lines were not shifted equally implying that classical physics was incorrect and his experiment is now used as one of the experimental tests in support of special relativity.

■ Length contraction

We have also seen that when an observer is moving relative to a length then that length will appear to be shortened. The longest possible length we called the proper length, L_0, and all other inertial observers will measure the length, L, between two events to be shorter than this.

$$L = \frac{L_0}{\gamma}$$

This equation is given in the *Physics data booklet*.

Derivation

This derivation is less obvious because it requires an understanding of how length is measured and defined. In spacetime, length is the distance between the positions of two events. In order to measure this length correctly in any given inertial reference frame, an observer must measure the position of both events that define the length simultaneously.

Note that this does not mean that the length being measured has to be a proper length – instead, it is as though they have to freeze the image (or take a photograph) of the length being measured and then take the measurements from this frozen image by standing in the middle of it and using reflecting simultaneous flashes to calculate its length. This means that a length measurement requires $\Delta t = 0$.

Let's imagine that we are trying to calculate the proper length, L_0, of an object *stationary in S′* but moving in S. The object is of known length, L, in inertial reference frame S.

$$\Delta x' = \gamma(\Delta x - v\Delta t)$$

$$L_0 = \gamma(L - 0)$$

$$L = \frac{L_0}{\gamma}$$

41 A rod is measured to have a proper length of exactly 1.00 m. How long would you measure it to be if it was to fly past you at 0.80*c*?

42 The time taken for the rod in question 41 to pass a fixed point in the laboratory is 2.5×10^{-9} s. What time interval would an observer travelling with the rod measure between the same two events?

43 Rosie flies through space and, according to Rosie, her height is 1.60 m. Rosie flies headfirst past an alien spaceship and the aliens measure her speed to be 0.80c.
- **a** How tall will the aliens on their spaceship measure Rosie to be?
- **b** Jeanina takes 6.1×10^{-9} s to fly past the same aliens at 0.90*c*. According to the aliens what time interval does it take Jeanina to fly past them?

44 In a space race a spaceship, measured to be 150 m long when stationary, is travelling at relativistic speeds when it crosses the finish line.
- **a** According to the spaceship it takes 7.7×10^{-7} s to cross the finishing line. How fast is it travelling in terms of *c*?
- **b** What time interval does the spaceship take to cross the finishing line according to an observer at the finishing line?
- **c** According to an observer at the finishing line, how long is the spaceship?
- **d** According to the observer at the finishing line, how fast is the spaceship travelling?

45 In the same race as question 44 a sleek space cruiser takes only 2.0×10^{-6} s to cross the finish line according to the race officials at the line. They measure the space cruiser to be 450 m long. How long is the space cruiser according to its sales brochure?

Tests of special relativity – the muon-decay experiment

A muon is an exotic particle that behaves just like an electron but is 200 times more massive. Their relatively large mass means that they are unstable and decay rapidly. They can be produced relatively easily in high-energy physics laboratories, so they can be studied easily. The standard method of timing their lifespan uses a block of scintillating material to stop a muon. This produces a tiny scintillation, or flash of light, as the muon's kinetic energy is converted into a photon. When the muon decays it emits a high-energy electron and a couple of neutrinos. The electron causes a second flash and the time interval between these two events provides the lifetime of the muon. A stationary muon in the lab has an average lifetime of 2.2×10^{-6} s.

Radioactive decay is a random process so, when we are measuring the decay of large numbers of particles, it is more normal to talk about half-life. Mathematically an average lifetime of 2.2×10^{-6} s is equivalent to a half-life of 1.5×10^{-6} s. The muon therefore provides us with a tiny clock that can be accelerated to relativistic speeds.

Muons are produced naturally in the Earth's atmosphere as a result of collisions between atmospheric particles and very high-energy cosmic radiation that is continually bombarding us. This occurs at around 10 km above the Earth's surface; the muons produced have an average speed of around 0.995*c*.

Let us consider three different options to see if they match what is observed experimentally.

Newtonian

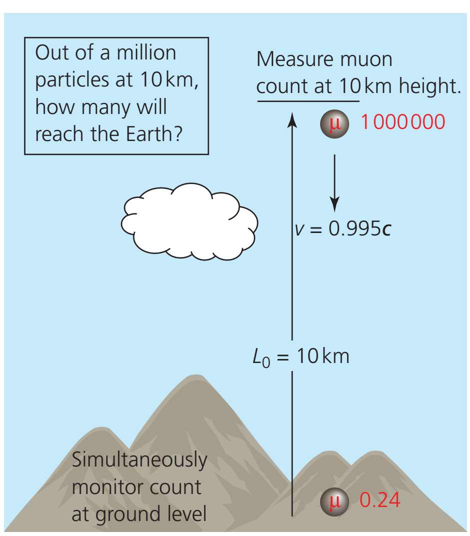
The time taken for the muons to reach the Earth's surface from a height of 10 km at a speed of 0.995*c* is:

$$t = \frac{x}{v} = \frac{10\,000}{0.995 \times 3.00 \times 10^8} = 3.35 \times 10^{-5}\,\text{s}$$

$$\text{number of half-lives} = \frac{\text{total time}}{t_{1/2}} = \frac{3.35 \times 10^{-5}}{1.5 \times 10^{-6}} = 22$$

$$\text{fraction reaching Earth's surface} = \left(\frac{1}{2}\right)^{22} = 2.4 \times 10^{-7}$$

Figure 13.17



Note that this is a vanishingly small fraction of the original number of muons at 10 km above the Earth's surface. If the Newtonian frame is correct, then almost no muons reach the Earth's surface because almost all would have decayed by that time (Figure 13.7).

Relativistic – Earth reference frame

In the Earth reference frame, the proper length is measured as 10.0 km and the speed of the muons is 0.995*c*. An observer in this frame of reference would also measure the time interval to be 3.35×10^{-5} s. However, they would know that an observer in the muon frame of reference would be measuring proper time, which is an invariant quantity, and can calculate this:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$= \frac{1}{\sqrt{1 - \frac{(0.995c)^2}{c^2}}}$$

$$= \frac{1}{\sqrt{1 - 0.995^2}}$$

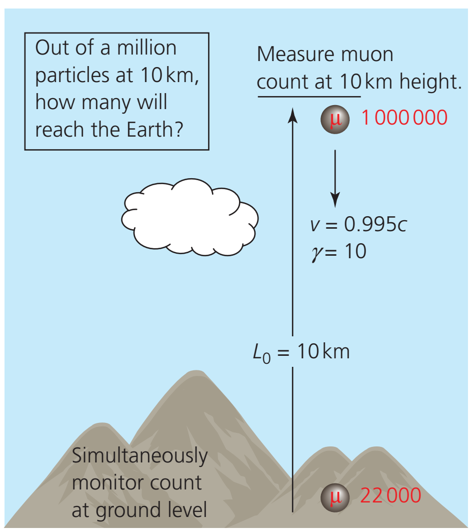
$$= 10.0$$

$$\Delta t_0 = \frac{\Delta t}{\gamma} = \frac{3.35 \times 10^{-5}}{10} = 3.35 \times 10^{-6}\,\text{s}$$

$$\text{number of half-lives} = \frac{\text{total time}}{t_{1/2}} = \frac{3.35 \times 10^{-6}}{1.5 \times 10^{-6}} = 2.2$$

$$\text{fraction remaining} = \left(\frac{1}{2}\right)^{22} = 0.22$$

■ **Figure 13.18**



This is about one fifth, so plenty of muons would actually reach Earth's surface compared with the number measured up at 10 km above ground. Confirmation of this result would therefore provide evidence in support of the idea of time dilation – the muons reach Earth because time has passed slowly for them, and so fewer have decayed than would have done in Newtonian physics (Figure 13.18).

Relativistic – muon reference frame

In the muon's frame of reference, the 10.0 km thickness of the lower atmosphere is contracted.

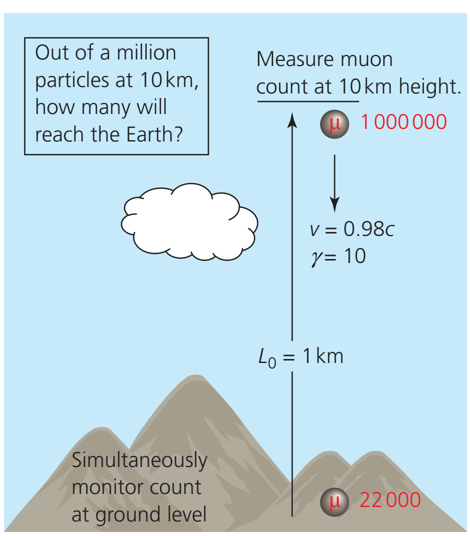
$$L = \frac{L_0}{\gamma} = \frac{10\,000\,\text{m}}{10} = 1\,000\,\text{m}$$

$$t = \frac{x}{v} = \frac{1\,000}{0.995 \times 3 \times 10^8} = 3.35 \times 10^{-6}\,\text{s}$$

So the fraction remaining to reach the Earth's surface is once again 0.22. This is because, from the muon's reference frame, the rest observer perceives what we measure to be 10 km of atmosphere to be only 1 km.

A confirmation of this result would therefore provide evidence in support of the idea of length contraction (Figure 13.19).

■ **Figure 13.19**



Experimental results

Does the experimental data support relativity? In general terms, a muon passes through every square centimetre of the Earth's surface every second, whereas 10 km higher the rate is approximately five times greater. This is convincing evidence for both the time dilation and length contraction aspects of special relativity. However, in reality the data are complicated – muons are produced in the atmosphere at a wide range of heights and with a very wide spectrum of different energies.

46 Some muons are generated in the Earth's atmosphere 8.00 km above the Earth's surface as a result of collisions between atmospheric molecules and cosmic rays. The muons that are created have an average speed of 0.99*c*.

- **a** Calculate the time it would take the muons to travel the 8.00 km through the Earth's atmosphere to detectors on the ground according to Newtonian physics.
- **b** Calculate the time it would take the muons to travel through the atmosphere according to a relativistic observer travelling with the muons.
- **c** Muons have a very short half-life. Explain how measurements of muon counts at an altitude of 8.0 km and at the Earth's surface can support the theory of special relativity.

13.3 (A3: Core) Spacetime diagrams – *spacetime diagrams are a very clear and illustrative way of showing graphically how different observers in relative motion to each other have measurements that differ from each other*

Spacetime was a concept first introduced by Minkowski in 1908. He was Einstein's former mathematics teacher. Einstein initially rejected the idea of spacetime but then realised its importance and used it as a major stepping stone in the discovery of general relativity. Spacetime diagrams, sometimes called Minkowski diagrams, can be a very powerful method of explaining relativistic physics. They contain a lot of information, so we will try to build up the pieces in several parts before putting a complete diagram together.

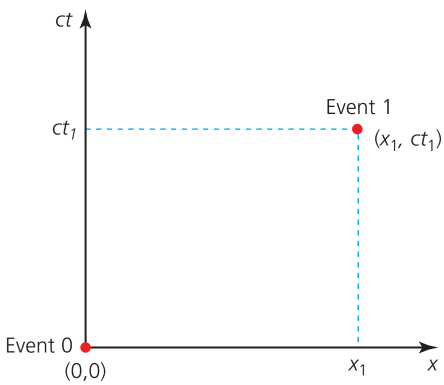


Figure 13.20 Spacetime diagram for an inertial reference frame, S, showing two events and their coordinates

Axes

Spacetime diagrams are normally drawn with the spatial dimension, *x*, on the horizontal axis and the temporal (i.e. time) dimension, *t*, on the vertical axis. Although the vertical axis could just show time, more commonly it shows the speed of light times time, *ct*, because this simplifies the diagram.

The axes represent the reference frame, or coordinate system, of a specific inertial observer.

Events

Events are represented as a point in spacetime. Just like an ordinary graph, the coordinates of the event are read off from the axes. In Figure 13.20 it is clear that, for an inertial observer in reference frame S, Event 0 occurs before Event 1.

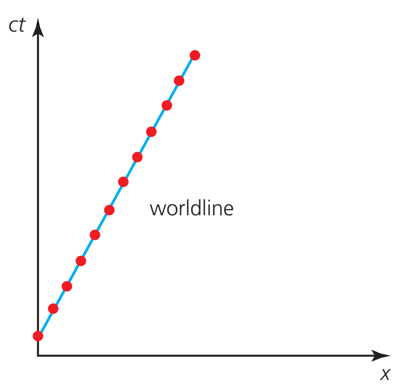


Figure 13.21 Spacetime diagram showing how a string of events joined together produces an object's worldline

Worldlines

An object travelling through spacetime can be imagined as a string of events. If we join up this line of events then we can plot an object's path through spacetime. We call this path the object's **worldline**. In Figure 13.21 a straight worldline is drawn showing that the object is moving through space with constant velocity relative to the observer.

The worldline shown does not pass through the origin because the object is observed a short time after the observer started their clock.

Gradient

The gradient of a worldline is given by c/v, so the steeper the gradient the slower the object is travelling. An object that is stationary as seen by an observer in this reference frame will have a vertical line because its x-coordinate does not change. The units are the same on each axis so that the gradient of 1 (i.e. a line drawn at 45°) represents the worldline (or **lightline**) of a photon through spacetime because $v = c$. All inertial observers agree on the value of c, so all observers must agree on the worldline for light as shown in Figure 13.22.

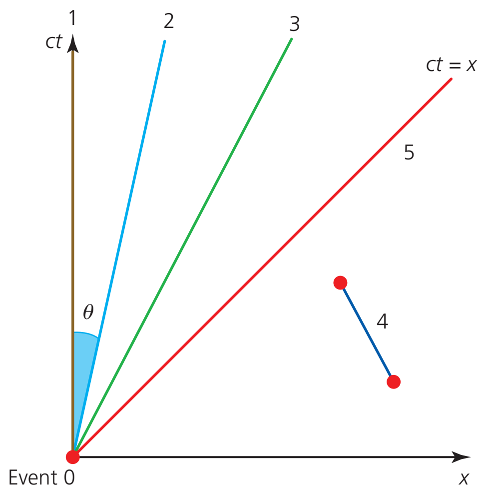


Figure 13.22 Spacetime diagram showing for inertial observer S:
1) worldline for a stationary object, this is the worldline of the observer in S;
2) worldline for a moving object;
3) worldline for faster moving object;
4) worldline for an object travelling in the opposite direction to the other objects between two events;
5) worldline for a photon, this is along the line where $ct = x$ and so has a gradient of 1.

The angle between worldlines

The angle between any straight worldline and the ct axis is given by:

$$\theta = \tan^{-1}\left(\frac{v}{c}\right)$$

This equation is given in the *Physics data booklet* and can be derived from Worked example 11.

Adding reference frame S′

Representing a second inertial reference frame on the same diagram is straightforward because the background spacetime does not change and events do not need to be moved, making it possible to compare how different observers perceive the same events.

Instead, the axes in the second reference frame become skewed. Suppose we look at the object that made worldline 2 in Figure 13.22. In Figure 13.23 the object is at rest in its own reference frame (S′) and so the ct' axis is along the object's worldline; the x' axis mirrors the ct' axis in the lightline.

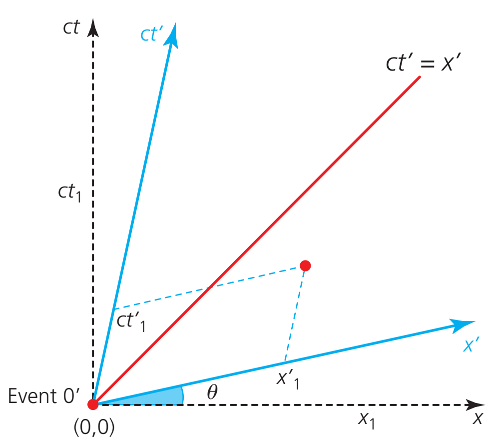


Figure 13.23 Spacetime diagram showing the additional axes for reference frame S′ in blue

According to an observer at rest in S′ the coordinates of the event are now (x'_1, ct'_1). To understand this we need to review the ideas of 'simultaneous' and 'stationary'.

In reference frame S′ events that are simultaneous occur along a line parallel to the x' axis, while events that occur in the same place occur along a line parallel to the ct' axis.

It is unlikely that you will have come across axes of this form before so spend some time thinking about how they work. This is further explained in Figure 13.24 and Figure 13.25.

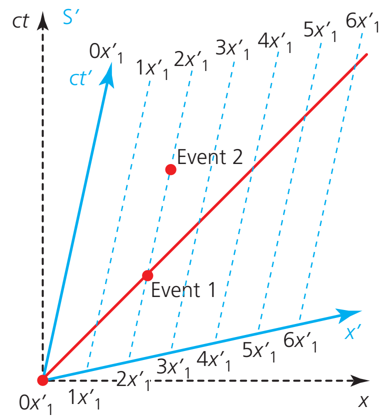


Figure 13.24 Spacetime diagram showing dashed blue lines that mark separate worldlines for points that are stationary in reference frame S′. The graph works in exactly the same way as the graphs you are used to, except that the grid is skewed rather than vertical and horizontal. Events 1 and 2 occur at the same position in space relative to an observer in S′

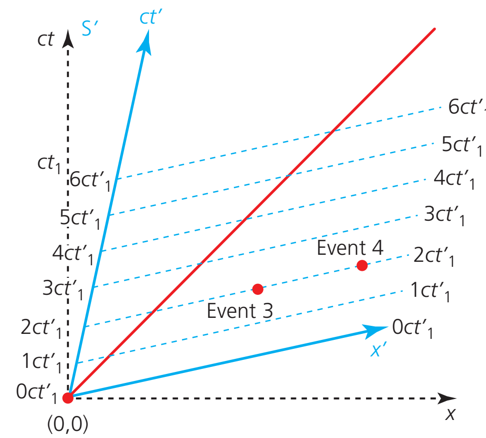


Figure 13.25 Spacetime diagram in which six lines parallel to the x' axis are marked. These lines join points that are simultaneous with each other, i.e. according to an observer in S′ all the events along line $1ct'$ occur simultaneously. At twice this time, all the events along dashed line $2ct'$ occur simultaneously, and so on. Events 3 and 4 are therefore simultaneous for an observer in S′

Worked example

11 Use the Lorentz transformation equations to show that the x' line has a gradient of $\frac{v}{c}$, and hence confirm that the angle between the x and x' axes is given by $\theta = \tan^{-1}\left(\frac{v}{c}\right)$, as shown in Figure 13.26.

The equation for the x' axis in terms of x and t can be found by setting $t' = 0$ and using the Lorentz transformation for t':

$$t' = \gamma\left(t - \frac{vx}{c^2}\right) = 0$$

so the bracket = 0.

$$t = \frac{vx}{c^2}$$

$$ct = \left(\frac{v}{c}\right)x$$

which is of the form $y = mx$.

$$\text{gradient} = \left(\frac{v}{c}\right) = \frac{c\Delta t}{\Delta x} = \frac{opp}{adj}$$

$$\tan\theta = \frac{opp}{adj} = \frac{v}{c}$$

$$\theta = \tan^{-1}\left(\frac{v}{c}\right) \text{ as required.}$$

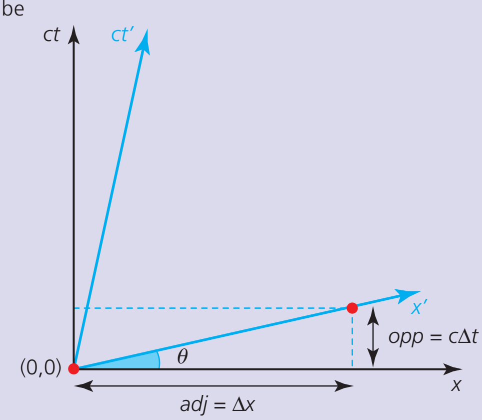


Figure 13.26 Spacetime diagram showing the calculation of the angle formula

Units on spacetime diagrams

Intriguingly, the units on the x and x' axes are not of the same size and this complicates how the geometry of spacetime can be used to solve problems. Remember that the spacetime interval, Δs^2, is an invariant quantity, so all observers will measure the same spacetime interval between two events no matter how fast the observers are travelling.

Let's consider four different inertial observers all travelling at different speeds, as shown in Figure 13.27. They are each asked to record a unit length that is stationary with respect to them. For each of them this is a proper length, of 1 unit, so each has a spacetime interval of $\Delta s^2 = -1$.

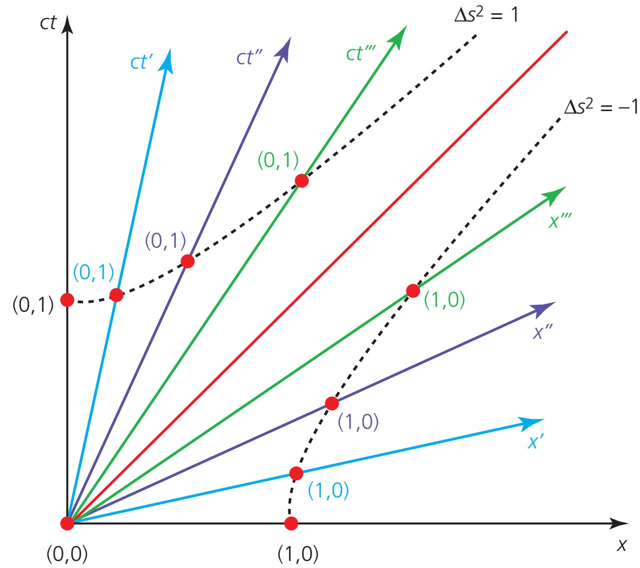


Figure 13.27 Spacetime diagram showing how the units vary for four different reference frames, coloured black, blue, purple and green

The black dashed lines link all the points with spacetime interval = 1 and −1, measured from the origin. They form a curve called a hyperbola. From this we can see that *the scale of 1 unit changes for each of the different x axes*. Although it is possible to calculate the exact coordinates of the hyperbola the mathematics to do this is beyond the IB syllabus so you can just stretch them on.

Similarly, the units on the time axes behave in a symmetrical fashion.

47 Use the Lorentz transformation equations to show that the *ct′* line has a gradient of *c/v*, and hence confirm that the angle between the *ct* and *ct′* axes is also given by:

$$\theta = \tan^{-1}\left(\frac{v}{c}\right)$$

48 a Use a ruler, calculator and the spacetime diagram in Figure 13.28 to complete the following table:

Event	Coordinates in S (*x*,*ct*)	Coordinates in S′ (*x′*,*ct′*)
A	(0,0)	(0,0)
B		(0.4,0.4)
C	(1.6,1.1)	(1,0)
D		
E		(0,1)
F		
G		

b List the order that the events occur in according to an observer in S and an observer in S′.

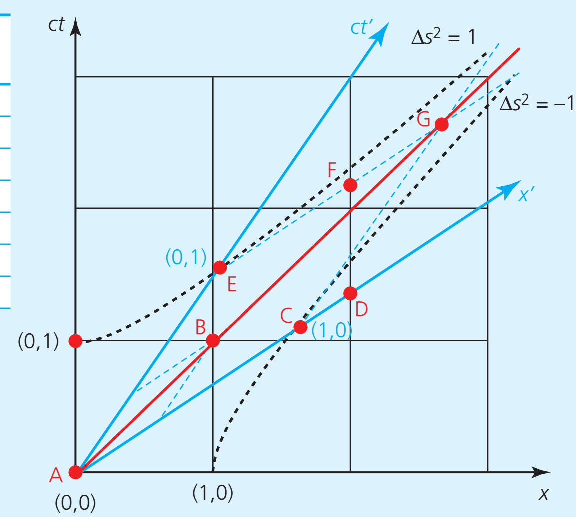
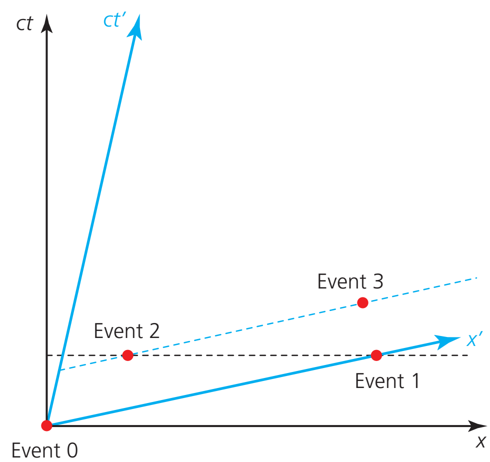


Figure 13.28 Spacetime diagram showing seven events, labelled A to G, from two different reference frames

Simultaneity in spacetime diagrams

Remember that all inertial observers will agree that two events are simultaneous if they occur in the same place, but they may disagree as to the order of the two events that occur at two different points in space. Figure 13.29 shows a spacetime diagram with four different events. According to one observer, Event 0 occurs first followed by Events 1 and 2 occurring simultaneously, with Event 3 happening last. However, for the other observer Events 0 and 1 both occur simultaneously followed by Events 2 and 3 occurring simultaneously.

Figure 13.29 Spacetime diagram comparing simultaneity in different reference frames



Worked example

12 Remember how an observer could demonstrate that two events were simultaneous (see Figure 13.9). Draw a spacetime diagram with reference frame S representing the observer on the train and S′ representing the reference frame on the platform.

The light rays are sent out in opposite directions, so we need to draw a positive and a negative x axis to allow us to position both the events (Figure 13.30).

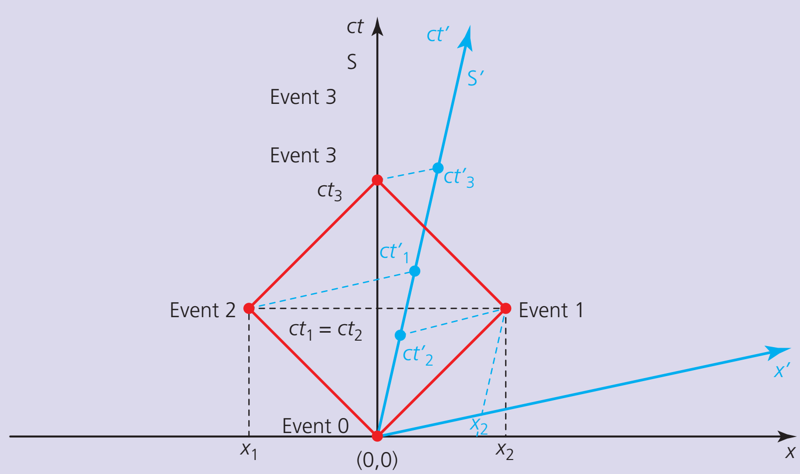


Figure 13.30 Spacetime diagram for the thought experiment defining simultaneous, carried out in Figure 13.9. The red lines represent the worldlines of the two reflecting light rays. The grey axes represent the inertial reference frame of the train carriage, S, while the blue axes represent the inertial frame of the platform, S′, with the train moving to the left. The dashed intersections with the timeline of each observer give their version of the order of events.

Observer S sees:

- the pulses are sent out simultaneously (Event 0)
- the pulses reach each end of the carriage simultaneously (Events 1 and 2)
- the pulses return to observer S simultaneously (Event 3).

Observer S′ sees:

- the pulses are sent out simultaneously (Event 0)
- the pulse that is fired down the carriage against the motion of the carriage must arrive at the end of the carriage before (Event 1) the pulse that is fired up the carriage (Event 2)
- however, S′ still sees the pulses return to observer S simultaneously (Event 3).

The geometry of the spacetime diagram gives us exactly the same outcome, demonstrating that events with no spacetime interval are simultaneous for all observers, but events that occur in two separate places can be simultaneous for some observers but not for others.

Nature of Science

Visualization of models

The visualization of the description of events in terms of spacetime diagrams is an enormous advance in understanding the concept of spacetime.

The Lorentz transformations we use to describe how we transfer from one reference frame to another are highly mathematical and make the topic very difficult to interpret in a non-mathematical way. The geometry of spacetime diagrams appears initially quite confusing, but with practice provides an entirely different way of approaching relativity. This new dimension means that aspects of relativity become significantly more accessible – in particular, spacetime diagrams readily explain whether or not events are simultaneous in different reference frames and explain the order of events seen by different observers.

With more practice, spacetime diagrams also explain concepts such as time dilation and length contraction but can also be used to understand relativistic velocity additions and to visualise why it is impossible to exceed the speed of light in a vacuum.

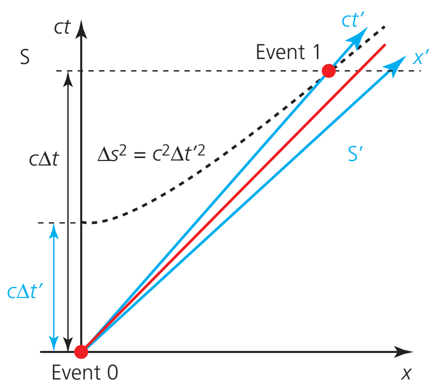


Figure 13.31 Spacetime diagram of the muon-decay experiment. Event 0 is the formation of a muon by the incoming cosmic radiation, while Event 1 is the arrival of the muon at the Earth's surface. The black reference frame, S, is the Earth reference frame, while the blue reference frame, S′, is that of the muon. An observer travelling with the muon measures the proper time between Events 0 and 1. To measure this on the scale of the vertical *ct* axis we follow the dashed line of constant spacetime interval from Event 1 to where it crosses the *ct* axis, where it can be easily calculated by measuring the interval labelled $c\Delta t'$. The interval according to an observer in reference frame S can be calculated by measuring $c\Delta t$

Time dilation in spacetime diagrams

Now look at a spacetime diagram for the muon experiment (Figure 13.31). Using the angle formula actually gives an angle of 44.9° for $v = 0.995c$, but ct' is drawn at less than this for clarity.

One of the problems with spacetime diagrams is that the scales on the axes are not the same. We could use Lorentz transformations to carefully mark the scales on each axis but there is a neat little trick that allows us to avoid this.

In the muon reference frame, Event 0 occurs a time $\Delta t'$ after Event 1, and both events occur at $x' = 0$. This is because in the muon reference frame the rest observer will see the stationary muon formed in the atmosphere (Event 0) and then the Earth's surface colliding with the stationary muon (Event 1) – the muon is stationary throughout. Thus the observer measures the spatial separation between the two events to be zero, so the time, $\Delta t'$, is the proper time between the two events, shown correctly to scale on the *ct* axis.

This occurs at a specific spacetime interval, where $\Delta s^2 = c^2\Delta t'^2$, and we can follow the dashed line that joins all the points with this same spacetime interval. Where this crosses the vertical *ct* axis it marks the equivalent interval as measured on the scale of the *ct* axis. On the spacetime diagram this is labelled $c\Delta t'$.

In the Earth reference frame, S, the time interval between Events 0 and 1 is significantly longer and can be found from the vertical coordinate of Event 1. This is marked as $c\Delta t$ on the *ct* axis. Since both measurements have been correctly scaled onto the *ct* axis their lengths can now be directly compared, and it is clear that the proper time interval $\Delta t'$ is considerably shorter than the streched (or dilated) Δt time interval. Showing that time has been stretched or **dilated**. Careful measurement from the *ct* axis would show that $c\Delta t = \gamma c\Delta t'$. Note that this equation appears to be the wrong way round because S′ is measuring proper time.

49 Use the spacetime diagram in Figure 13.32 and the Lorentz transformation equations to calculate:

a The velocity of an object that has a worldline at 30°.

b Calculate the value of γ for this speed.

c The time, t, at which an observer in reference frame S will record the object to have travelled 10 ly (Event 1) at this speed.

d The value of $c\Delta t$ between Event 0 and Event 1.

e The graph is drawn correctly to scale. Measure the length of $c\Delta t$ and $c\Delta t'$ on the ct axis and show that the ratio of the measured lengths $c\Delta t/c\Delta t' \approx \gamma$.

f State which reference frame is measuring proper time.

g Use time dilation to calculate the value of $c\Delta t'$.

h Mark the position of 14 ly on both the vertical black axes (S reference frame) and blue axes (S′ reference frame) to show that the scales on the axes are different.

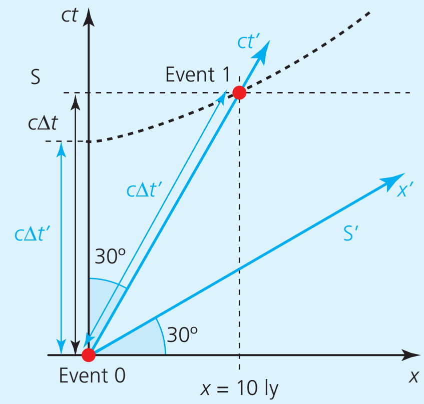


Figure 13.32

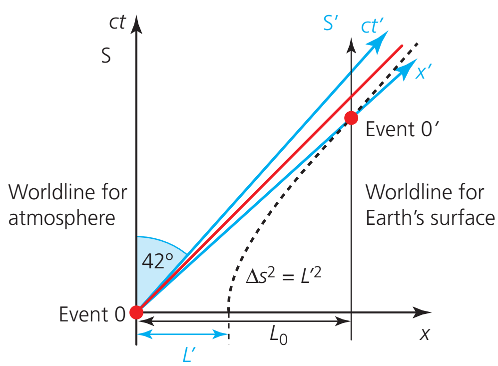


Figure 13.33 Spacetime diagram for the muon-decay experiment used to demonstrate length contraction. The instantaneous separation between the atmosphere and the Earth's surface in the muon reference frame must be measured in each reference frame

Length contraction in spacetime diagrams

Once again let us turn to the muon experiment as shown in Figure 13.33. The length being measured is the distance between the formation of the muons in the Earth's atmosphere and the surface of the Earth.

In the previous section, the length contraction equation was harder to derive than the time-dilation equation because it required one more key piece of information– in order to measure a length correctly we must measure the position of each end of the length at the same time. In other words, the two spacetime events used to determine the two ends of the length in a given reference frame must occur simultaneously in that reference frame.

This length is straightforward to measure in the Earth reference frame because it is simply the horizontal separation between the vertical worldline of the Earth's atmosphere and the worldline of the Earth's surface. These are shown on the spacetime diagram as the two vertical black axes. Because each is stationary in the Earth reference frame, the separation between them is a proper length and is labelled L_0 on the diagram, where it could easily be measured on the x axis.

In the muon reference frame, S′, the distance between the Earth's atmosphere and the Earth's surface can be measured using simultaneous events 0 and 0′, where the worldline of the atmosphere and the worldline of the Earth's surface each cross the x' axis. In the muon reference frame both events occur when $t' = 0$, so they can be used to correctly measure the separation, L' – it could be measured off the scale on the x' axis but we would need to calculate the scale to do this.

Instead, we can sketch on the curve that links all the points with spacetime interval $\Delta s^2 = -L'^2$. Extending this to the x axis gives the separation between Events 0 and 0′ on the x axis scale, where it can be easily be measured. It can clearly be seen that the proper length is much larger than the contracted length, L', confirming that length contraction occurs. Careful measurement would also show that $L' = L_0/\gamma$. Once again the spacetime diagram's geometry has been able to represent the dynamics of relativity.

50 In Figure 13.33 the angle between the x and x' axes is also 42°. Calculate:

a The relative velocity of the two frames of reference.

b The value of γ.

c If $L_0 = 1.0\,\text{m}$ calculate the length of L'.

d Using a ruler, measure the ratio of L_0/L' to confirm that this gives the value of γ.

51 Einstein's first postulate stated that the laws of physics are the same in all inertial reference frames. This means that we should be able to show on a spacetime diagram that an object that is stationary in reference frame S′ will also be measured as having a contracted length by an observer in S. Use Figure 13.34 to show that this is indeed the case by marking on the length as measured by S′ on the x' axis and using the spacetime interval curve to mark on the x' axis the equivalent length as measured by S. Hence, use the measured lengths along the x' axis to estimate the value for γ.

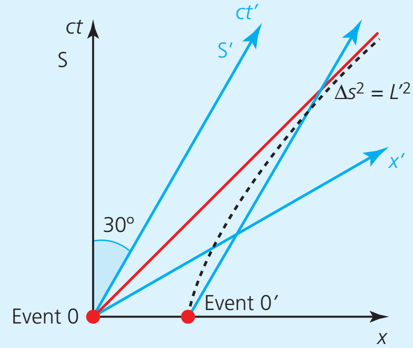


Figure 13.34

The twin paradox

One of the early concerns raised by special relativity was its apparent symmetry. If two inertial observers each carrying their own metre-long rod pass each other then they would each see the other's rod being contracted, and therefore shorter than their own. Similarly, they can each read the other's clock as running slow relative to their own. The rules of special relativity are not broken here because these situations are symmetrical. This is discussed in Figure 13.35.

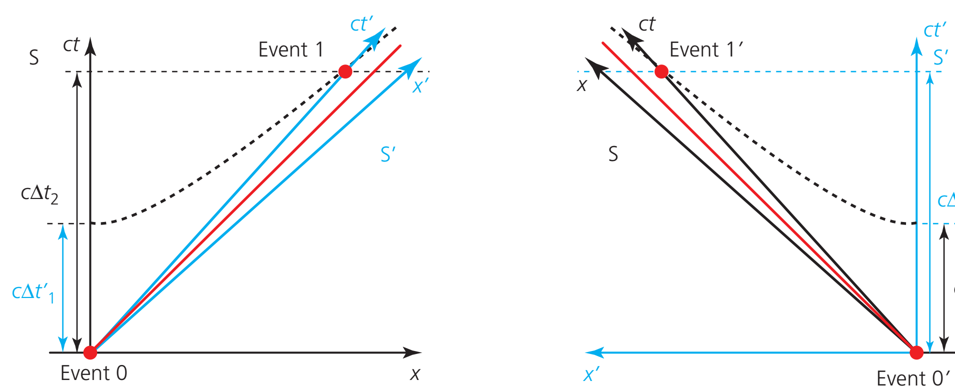


Figure 13.35 The symmetry of relativity means that if two observers are carrying out the same experiment on each other the results should be symmetrical. Suppose two observers are travelling in opposite directions. They each carry a metre-long rod held parallel to their line of motion and two stopwatches. They each time two different intervals. Firstly, they start timing when they meet the front of the other observer's rod and stop timing when they meet the back of the other's rod, Δt_1 and $\Delta t'_1$; secondly, they record the time interval for the other observer to pass the front and back of their own rod, Δt_2 and $\Delta t'_2$. According to relativity, the results must be symmetrical and the spacetime diagrams confirm this. Each observer correctly measures a proper time (Δt_1 or $\Delta t'_1$) when they pass the ends of the other observer's rod, but measure a dilated (stretched) time when they record the other observer passing their own rod (Δt_2 or $\Delta t'_2$). This only works because Events 0 and 0′, and 1 and 1′ are four different events and means that the two spacetime diagrams are mirror images of each other

Indeed the postulates of special relativity clearly state that these inertial reference frames have equal validity and so both versions must be correct.

The twin paradox asks what would happen if we had two twins, Andrew and Alice – Andrew likes to stay at home so he remains on Earth; Alice likes to travel so she is sent to our nearest neighbouring star in a spaceship that travels at close to the speed of light. When Alice arrives on Proxima Centauri she decides she does not like the double sunrise of the binary system and returns home in a different ship travelling in the opposite direction at the same speed. According to relativity Andrew must think that Alice will not age as much as he has because she has travelled. In Alice's reference frame though, she is stationary and she sees the Earth shoot away from her spaceship; she then jumps on a second spaceship and the Earth comes rushing back towards her, so she can equally argue that Andrew has travelled and she has not, so he will have aged less than her. The paradox is that they cannot both be younger than the other, and yet the situation appears symmetrical.

Solution to the twin paradox

The solution is that the twin paradox is not symmetrical. Alice cannot argue that she has not travelled because she has experienced a dramatic acceleration when she turned round to come back (breaking the requirements for special relativity). This means that she has not been in one inertial reference frame but in two different inertial reference frames. Figure 13.36 shows the lack of symmetry when spacetime diagrams are compared. Reference frame A is from Andrew's perspective while A′ is that of the spaceship that first took Alice to Proxima Centauri. The purple worldline *ct″* is that of the second spaceship that took Alice back to Earth.

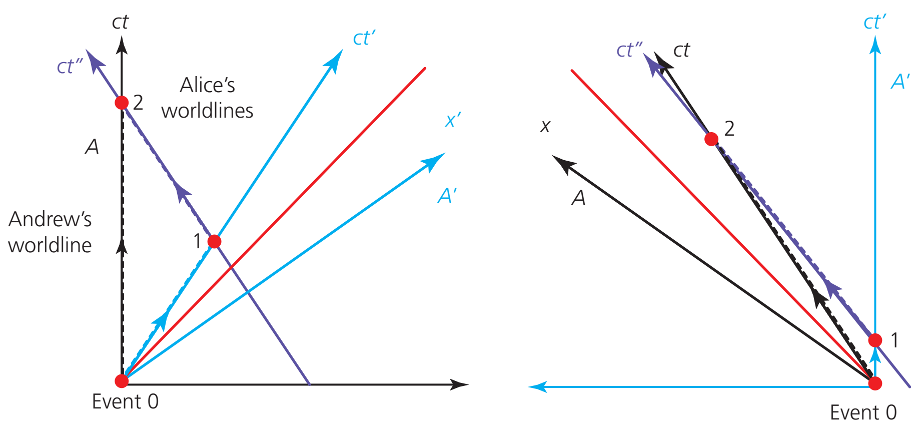


Figure 13.36 Resolving the twin paradox. Spacetime diagrams from Andrew's perspective and from the perspective of the first half of Alice's motion clearly show that Alice's and Andrew's situations are not symmetrical, as required by the paradox, so the paradox is resolved. The black *ct* and *x* axes are for Andrew, the blue *ct′* and *x′* axes are for the first half of Alice's trip and the purple *ct″* axis is for the second half of her trip. Andrew's worldline between Events 0 and 2 is marked by the black dashed lines; Alice's worldline between 0 and 1 is marked by the blue dashed line and between 1 and 2 by the purple dashed line

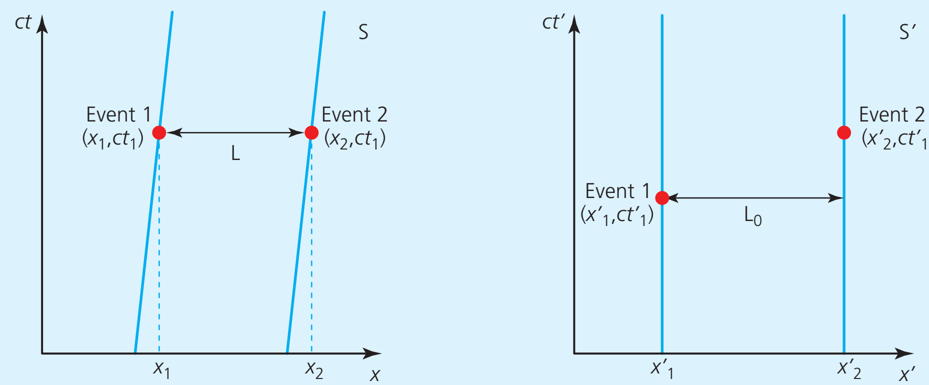
Utilizations

The Hafele–Keating experiment

In 1971 a version of the twin paradox experiment was carried out by Joseph Hafele and Richard Keating (Figure 13.37). Alice and Andrew were not people, instead Hafele and Keating used four very precise atomic clocks, Andrew (represented by atomic clocks that stayed at the US Naval Observatory) stayed on Earth while Alice went on two commercial plane journeys, represented by clocks that were flown right around the Earth. One clock flew eastwards with the rotation of the Earth and one clock flew westwards against the Earth's rotation. When the atomic clocks returned they were compared with the Earth-bound clocks.

The results were published in the journal *Science* and confirmed the predictions of relativity. The eastbound clock had run −59 ± 10 ns slow while the westbound clock had run 273 ± 7 ns fast, and within the uncertainties predicted matched the expectations of relativity. The two values do not add up to zero because there is also an effect caused by the planes flying in a weaker gravitational field, causing both clocks to run faster. This is explained in the section on general relativity.

52 Look at the two spacetime diagrams in Figure 13.37. The blue worldlines represent each end of a rod. An observer in reference frame S sees the rod moving past at constant velocity, while the observer in reference frame S′ is at rest with the rod. Both observers measure the length of the rod using two events, one at each end of the rod.



■ **Figure 13.37**

a Explain why the observer in reference frame S can only correctly measure the length of the rod if the two events occur simultaneously.
b Describe how this might be achieved experimentally.
c Explain why the observer in reference frame S′ does not need simultaneous events to correctly measure the length of the rod.
d The two observers measure different values for the length of the rod. Which observer is measuring the proper length? Explain your answer.
e What aspect of special relativity do the two spacetime diagrams demonstrate? Explain your answer.

53 The 1971 experiment in which four atomic clocks are compared after two have been on a fast flight does not exactly match the twin paradox that has been described. Describe and explain how the two are different.

ToK Link

Can paradoxes be solved by reason alone, or do they require the utilization of other ways of knowing?

The twin paradox is an example of a paradox that occurs because of a misunderstanding in the interpretation of what is known. This depends on the exact definition of paradox being used, so either it is not a true paradox or some paradoxes can be resolved through more careful reasoning.

13.4 (A4: Additional Higher) Relativistic mechanics – *energy must be conserved under all circumstances, and so must momentum; the relativity of space and time requires new definitions for energy and momentum in order to preserve the conserved nature of these laws under relativistic transformations*

■ Rest mass and electrical charge as invariant quantities

In Sections 13.1 to 13.3 we discussed several quantities that do not vary, no matter how fast an object was perceived to be travelling. These were the spacetime interval, Δs^2, the proper time, t_0, and the proper length, L_0. In addition, all observers will agree on the rest mass, m_0, and electrical charge, q.

We can think of the rest mass as being the 'proper mass' of an object. It is defined as the mass as measured by an observer who is stationary relative to the object. The relationship between rest mass and the measured mass, m, is given by:

$$m = \gamma m_0$$

This equation is not given in the *Physics data booklet*.

Worked example

13 A particle is measured to be travelling at exactly 0.85c and has an inertial mass of 1.9149 u. Suggest what type of particle it might be.

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$\gamma = \frac{1}{\sqrt{1 - 0.85^2}} = 1.8983$$

$$m = \gamma m_0$$

$$m_0 = \frac{m}{\gamma}$$

$$= \frac{1.9149\,\text{u}}{1.8983} = 1.0087\,\text{u}$$

Looking at the *Physics data booklet*, this suggests that the particle could be a neutron.

Total energy and rest energy

If velocity is a relative quantity, no longer with an absolute zero but with an absolute maximum, then an object's energy must also depend on the reference frame from which it is viewed. Remember that:

total energy = potential energy + kinetic energy

If an object is stationary within a reference frame then it has no kinetic energy and therefore it has its lowest energy. We call this the **rest energy**, E_0. As an object moves faster it gains kinetic energy but, fundamentally, it can never reach the speed of light in a vacuum. Relativistically, the reason for this is because an observer travelling in the reference frame of the object must still measure the speed of light in a vacuum to be the invariable constant, c, and this is not the case if the object reaches or exceeds the speed of light itself.

So, how does the physics explain this? The answer comes from the equation:

$$E = mc^2$$

The rest energy is given by:

$$E_0 = m_0c^2$$

This equation is given in the *Physics data booklet*.

This is different from Newtonian physics, in which mass is an invariant quantity, so at speeds approaching the speed of light in a vacuum the classical kinetic energy equation $\left(\frac{1}{2}m_0v^2\right)$ fails and we must use a relativistic kinetic energy equation. What this tells us is that mass and energy are equivalent. As discussed in Chapter 7, if an object has energy then it has mass; if an object gains more energy then it gains more mass. As an object accelerates it gains more energy and therefore gains more mass, and becomes increasingly difficult to accelerate further.

The total energy of an object can be easily calculated by multiplying the rest energy by the Lorentz factor:

$$E = \gamma m_0c^2$$

This equation is given in the *Physics data booklet*.

From this, and the equation for the rest energy, we can calculate the *relativistic* kinetic energy:

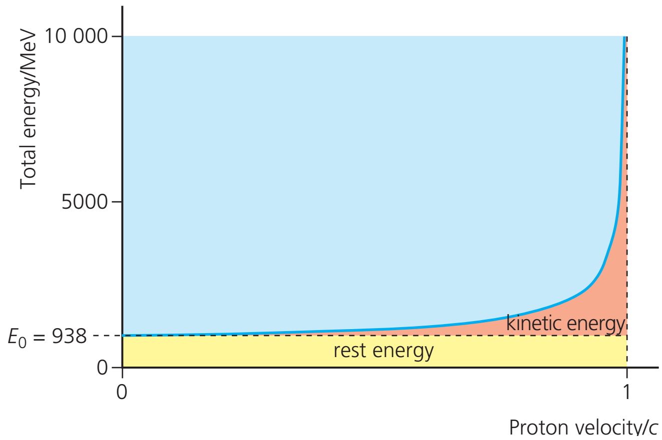
relativistic kinetic energy = total energy – rest energy

$$E_k = \gamma m_0c^2 - m_0c^2$$

$$= (\gamma - 1)m_0c^2$$

This equation is given in the *Physics data booklet*.

An example graph of the total energy of a particle is shown in Figure 13.38.



■ **Figure 13.38** Graph of a proton's total energy. The key things to recognise on the graph are: i) the rest energy is the constant value of E_0; ii) the relativistic kinetic energy is the difference between the total energy line and the rest energy; and iii) the proton's velocity never quite reaches c. Since mass and energy are equivalent, a similar shaped graph can be drawn for mass, with the rest mass replacing the rest energy and the units changing from MeV to MeV c^{-2}

Worked examples

14 A proton is measured according to the laboratory's reference frame to be travelling at 0.500c. According to observers in the laboratory, what is the proton's total energy?

$$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$$

$$= \frac{1}{\sqrt{1-0.500^2}} = 1.15$$

$$E = \gamma m_0 c^2$$

$$= 1.15 \times 938\,\text{MeV}\,c^{-2} \times c^2$$

$$= 1.08\,\text{GeV}$$

15 An electron is measured according to the laboratory's reference frame to be travelling at 0.75c. According to observers in the laboratory, what is the electron's kinetic energy?

$$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$$

$$= \frac{1}{\sqrt{1-0.75^2}} = 1.51$$

$$E = (\gamma - 1) m_0 c^2$$

$$= (1.51 - 1) \times 0.511\,\text{MeV}\,c^{-2} \times c^2$$

$$= 0.26\,\text{MeV}$$

Nature of Science

Paradigm shift

Einstein realized that the law of conservation of momentum could not be maintained as a law of physics. He therefore deduced that in order for momentum to be conserved under all conditions, the definition of momentum had to change and along with it the definitions of other mechanics quantities such as kinetic energy and total energy of a particle. This was a major paradigm shift.

When dramatic changes are made to the laws of science it is important to be able to differentiate between those aspects of the previous theory that are sound and those aspects that need to be changed. In this case Einstein realised that the core rules of conservation of energy and conservation of momentum needed to be upheld in special relativity and therefore the classical equations, which for physics students may wrongly appear to be fundamental definitions, need to be revised.

In 1911 Einstein considered a variable-speed-of-light theory in which the speed of light in a vacuum is still an unbreakable speed limit as described by special relativity, but this speed limit changes in a changing gravitational field. He eventually abandoned this line of research. However, subsequent scientists have proposed it as a solution to various problems in cosmology. Most recently, Andreas Albrecht and João Magueijo suggested that if the speed of light were 60 times faster in the early universe then this would solve an issue in cosmology called the horizon problem and provide an alternative to mainstream inflation theory. What are the potential implications of changing the value of *c*?

Relativistic momentum

Because both the energy, and therefore the mass of a particle, increase with velocity, this must also have implications for calculating the momentum of particles travelling at speeds approaching *c*. Because the mass is directly linked to the rest mass by the Lorentz factor, the same must be true of relativistic momentum:

$$p = \gamma m_0 v$$

This equation is given in the *Physics data booklet*.

From this the English physicist Paul Dirac derived the relativistic equation for a particle's total energy:

$$E^2 = p^2c^2 + m_0^2c^4$$

This equation is given in the *Physics data booklet*.

This equation looks considerably more complicated to use so it is worth practising.

Worked example

16 An inertial observer measures a proton to be travelling with a velocity of 0.75c. Calculate both the proton's momentum and total energy according to the inertial observer.

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$= \frac{1}{\sqrt{1 - 0.75^2}} = 1.51$$

$$p = \gamma m_0 v$$

$$= 1.51 \times 938\,\text{MeV}\,c^{-2} \times 0.75\,c$$

$$= 1.06 \times 10^3\,\text{MeV}\,c^{-1}$$

$$E^2 = p^2c^2 + m_0^2c^4$$

$$E = \sqrt{p^2c^2 + m_0^2c^4}$$

$$= \sqrt{(1.06 \times 10^3\,\text{MeV}\,c^{-1})^2c^2 + (938\,\text{MeV}\,c^{-2})^2c^4}$$

$$= \sqrt{(1.06 \times 10^3)^2 + 938^2} \text{ because all the cs cancel out}$$

$$= 1.420 \times 10^3\,\text{MeV}$$

The units of energy, mass and momentum

You should already be familiar with the electronvolt as a unit of energy. More commonly particle energies are quoted in kiloelectronvolts (keV), megaelectronvolts (MeV) and gigaelectronvolts (GeV). The definition of the electronvolt comes from the equation for electrical potential energy, $E = qV$.

From Einstein's equation ($E = mc^2$) we know that:

$$m = \frac{E}{c^2}$$

The units of mass are eV c^{-2}, but more commonly these will be quoted as MeV c^{-2} or GeV c^{-2}.

Looking at Worked example 16 you can hopefully see that it also makes sense to introduce a new unit for momentum. This greatly simplifies the equations when the velocities are also given as fractions of c and, although it takes a bit of getting used to, just like the eV unit for energy, with practice the benefits soon become apparent.

$$p = mv$$

The units of momentum are eV c^{-1}, but more commonly these will also be quoted as MeV c^{-1} or GeV c^{-1}.

54 The benefit of the relativitistic units for energy, mass and momentum is that they make relativistic equations involving c much simpler. Fortunately, we do not normally have to convert into SI units but you should be able to do this.

a Convert 1.00 eV to joules.
b Convert 1.00 J into eV.
c How much energy does a mass of 1.00 kg possess?
d Convert 1.0 kg into eV c^{-2}.
e Convert 1.0 kg m s^{-1} into eV c^{-1}.

Utilizations

Nuclear power, particle accelerators and particle detectors

The laws of relativistic mechanics are needed to understand the energy and momentum changes that occur in nuclear power plants, particle accelerators and particle detectors.

Nuclear power plants

Nuclear power plants use a fission reaction to release some of the energy stored in unstable, large nuclei such as uranium-235 and plutonium-239. Spontaneous uranium-235 fission typically results in the release of 200 MeV of energy, mostly as the kinetic energy of the product nuclei due to the coulomb repulsion. An average of 2.5 neutrons with typical kinetic energies of 2 MeV are released along with gamma rays with a total energy of around 7 MeV. These neutrons are then slowed in collisions with a moderator material, such as heavy water, to kinetic energies of around 10 eV. At these energies the neutrons can be absorbed by other uranium-235 nuclei. Without understanding the implications of relativity, harnessing nuclear energy would not be possible.

Particle accelerators

Particle accelerators have become increasingly powerful. They have allowed us to explore the extremely high-energy densities of the early universe and to create a wide range of exotic new particles. However, they also have many industrial and even domestic uses – for example producing electron beams for irradiation, acting as neutron generators, and using proton beams to produce proton-rich ions (as opposed to the neutron-rich ions produced in fission reactors). Even old style cathode-ray television sets are a basic form of particle accelerator.

Modern research particle accelerators tend to be of two forms: linear accelerators that have a fixed target and circular accelerators that circle two beams in opposite directions. The advantage of a linear accelerator is that it is much cheaper to build – however, it is limited in the energies it can achieve by the length of the accelerator. Circular accelerators, such as synchrotrons, get round this problem by sending the particles repeatedly around a ring. The particle beam goes through dipole magnets that bend the beam inwards and quadropole magnets that focus the particle beam back into a tight cluster. The maximum energy of a synchrotron is therefore determined by both the radius of the accelerator ring and the strength of the magnetic dipoles that bend the particle path into a ring. However, both linear and circular accelerators pass the beam through electrical plates connected by oscillating high-energy electric fields, which cause the particles to travel faster.

Particle detectors

The job of a particle detector is to track and measure the energy, momentum and charge of the particles produced in high-energy particle collisions. Large magnetic fields are applied that result in a curvature of the path of the charged particles. Tracking the particles is done by a variety of different methods. Early detectors used photographic film, and bubble or cloud chambers. These were developed into spark chambers and scintillation chambers in which flashes of light are recorded on photographic film. Modern detectors now use solid-state semiconductors to track particles with very high precision. They are also able to cope with the huge number of particle tracks that are produced in modern particle accelerators.

55 1.00 kg of uranium-235 contains approximately 2.56×10^{21} atoms of uranium. A modern nuclear reactor replaces approximately 25 tonnes of uranium-235 every 2 years. If each uranium-235 fission results in 200 MeV of energy being released, what is the theoretical maximum power output of the power station? (2 years = 6.31×10^7 s)

56 Use conservation of momentum to explain why more energy is available to produce new particles in a synchrotron where two particle beams travelling in opposite directions collide, as opposed to a linear accelerator that collides accelerated particles into a fixed target, even if the total collision energy of the two accelerators is the same.

Particle accelerators

A particle accelerator works by using an electrical field to give kinetic energy to the particles (Chapter 10). Particle accelerators must of course obey the law of conservation of energy but, at speeds approaching the speed of light in a vacuum, we need to apply the relativistic energy equations. All the energy supplied by the field must be turned into relativistic kinetic energy:

loss of electrical potential energy = gain in kinetic energy

$$qV = \Delta E_k$$

This equation is given in the *Physics data booklet*.

Worked examples

17 An electron is accelerated from rest by a potential difference of 10 kV. Calculate the electron's velocity after this acceleration.

$qV = \Delta E_k = E_k$, since there is no initial kinetic energy

$$qV = (\gamma - 1)m_0c^2$$

$$\gamma = \frac{qV}{m_0c^2} + 1$$

$$= \frac{10 \times 10^3\,\text{eV}}{0.511 \times 10^6\,\text{eV}} + 1$$

$$= 1.02$$

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$\frac{v}{c} = \sqrt{1 - \frac{1}{\gamma^2}}$$

$$= \sqrt{1 - \frac{1}{1.02^2}}$$

$$= 0.20$$

$$v = 0.20c$$

18 According to Newtonian physics, what voltage is required to accelerate an electron from stationary to the speed of light, and what is the actual speed of the electron after taking relativistic effects into consideration? (Rest mass of the electron, $m_0 = 0.511\,\text{MeV}\,c^{-2}$)

$qV = \Delta E_k = E_k$, because there is no initial kinetic energy

$$qV = \frac{1}{2}m_0 v^2$$

$$V = \frac{m_0 v^2}{2q}$$

$$= \frac{0.511 \times 10^6\,\text{eV}}{2e}$$, because the c^{-2} and the c^2 cancel out

$$= 256\,\text{kV}$$

$$qV = (\gamma - 1)m_0 c^2$$

$$\gamma = \frac{qV}{m_0 c^2} + 1$$

$$= \frac{256 \times 10^3\,\text{eV}}{0.511 \times 10^6\,\text{eV}} + 1$$

$$= 1.50$$

$$\frac{v}{c} = \sqrt{1 - \frac{1}{\gamma^2}}$$

$$= \sqrt{1 - \frac{1}{1.50^2}}$$

$$= 0.75$$

$$v = 0.75c$$

57 In a laboratory an electron is timed as it passes across a 5.00 m gap after it has been accelerated from rest across a voltage of 500 kV.
  a What is the relativistic velocity of the electron?
  b What time interval should it take to cross the gap according to an observer in the laboratory frame of reference?
  c What time interval should it take to cross the gap according to an observer in the electron's rest frame of reference?

58 The large hadron collider at CERN should reach proton energies of 7 TeV (7×10^{12} eV) after its upgrade in 2015. How fast is a proton travelling if it has exactly 7 TeV of kinetic energy?

59 The muons used as tiny clocks in the atmospheric muon-decay experiments actually have a range of speeds. What is the kinetic energy of a muon travelling at 0.995c? The rest mass of a muon is 207 times bigger than the rest mass of an electron.

60 The second most powerful particle accelerator ever built was the Tevatron at Fermilab, which ran from 1983 to 2011. It accelerated protons and anti-protons around a ring in opposite directions to speeds of 0.999954c. What is the total mass, kinetic energy and momentum of a proton travelling at this speed?

61 Neutrons produced in nuclear fission reactors have an average kinetic energy of 2.0 MeV. In order to initiate another fission reaction they must be slowed to only 10 eV through multiple collisions. What is the decrease in momentum of the neutrons?

62 Alpha particles have a typical kinetic energy of 5.000 MeV and a rest mass of 3.727 GeV c^{-2}. Calculate the relativistic velocity and momentum of a typical alpha particle.

63 The theoretical upper limit for the energy of cosmic rays that have travelled any significant distance through space is called the Greisen–Zatsepin–Kuzmin limit. This limits cosmic ray energies to 5×10^{19} eV. What is the momentum of a proton travelling with this total energy?

Photons

The only particles that can travel at the speed of light in a vacuum are photons – the packets of energy and momentum that travel as oscillations of electromagnetic fields. The highest energy photons can be formed when unstable high-energy particles decay, converting all, or some, of their energy into electromagnetic radiation.

In Chapter 7 you studied the relationship between photon energy and either their frequency or, if they are travelling through the vacuum, their wavelength. This is given by:

$$E = hf = \frac{hc}{\lambda}$$

Where h is Planck's constant = 6.63×10^{-34} J s.

Photons also have momentum, which is confusing because they do not have any rest mass. As a result, the relativistic momentum equation ($E^2 = p^2c^2 + m_0{}^2c^4$) becomes:

$$E^2 = p^2c^2 + 0$$

$$p = \frac{E}{c}$$

Using the equation for photon energy it is possible to link photon momentum to photon frequency and wavelength. This is given by the equation:

$$p = \frac{E}{c} = \frac{hf}{c} = \frac{h}{\lambda}$$

This equation linking photon energy and momentum is not given in the *Physics data booklet*.

Worked example

19 A pion, or pi meson, is a type of particle formed from an up quark or a down quark in combination with an up or down antiquark. Pions can be either be positive, neutral or negative depending on the exact combination, and have the symbol π^+, π^0 and π^- respectively. The neutral pion has a rest mass of 135.0 MeV c^{-2} but it is unstable, decaying into two gamma photons. Because both energy and momentum must be conserved, in the rest frame of the pion the two photons must have equal energies and travel in opposite directions. What is the wavelength of the gamma rays produced?

To solve this we need to convert the rest mass to its equivalent energy in joules (by multiplying the rest mass by $\times 10^6$ and dividing by $(3.00 \times 10^8)^2$ to convert MeVc^{-2} to eV, than multiplying by 1.50×10^{-19} C to convert eV to J).

$$E = \frac{hc}{\lambda}$$

$$\lambda = \frac{hc}{E}$$

$$= \frac{6.63 \times 10^{-34} \times 3.00 \times 10^8}{135.0 \times 10^6 \times 1.60 \times 10^{-19}}$$

$$= 9.21 \times 10^{-15} \text{ m}$$

20 In another experiment, a pion is measured to be travelling with a total energy of 500.0 MeV immediately before it decays. The two photons produced in the decay must conserve both energy and momentum and, in this instance, one travels forward along the pion's path and the other travels backwards. What is the wavelength of each photon?

First calculate the momentum of the muon:

$$E^2 = p^2c^2 + m_0^2c^4$$

$$p = \sqrt{\frac{E^2}{c^2} - m_0^2c^2}$$

$$= \sqrt{500.0^2 - 135.0^2}$$

$$= 481.4\,\text{MeV}\,c^{-1}$$

We will use subscripts 1 and 2 to differentiate the two photons. Now use both conservation of energy and conservation of momentum to derive two equations:

$$p_1 - p_2 = 481.4\,\text{MeV}\,c^{-1}$$

$$E_1 + E_2 = 500.0\,\text{MeV}$$

However, for a photon $E = pc$, so if we multiply all of the top equation by c then we get two equations that we can solve simultaneously by subtracting one from the other:

$$E_1 - E_2 = 481.4\,\text{MeV}$$

$$E_1 + E_2 = 500.0\,\text{MeV}$$

$$E_1 = 490.7\,\text{MeV}$$

$$E_2 = 9.3\,\text{MeV}$$

From these it is a simple matter to work out the wavelength of each photon, remembering to convert the energy from MeV into J:

$$E = \frac{hc}{\lambda}$$

$$\lambda_1 = \frac{hc}{E_1}$$

$$= \frac{6.63 \times 10^{-34} \times 3.00 \times 10^8}{490.7 \times 10^6 \times 1.60 \times 10^{-19}}$$

$$= 2.53 \times 10^{-15}\,\text{m}$$

and

$$\lambda_2 = \frac{hc}{E_2}$$

$$= \frac{6.63 \times 10^{-34} \times 3.00 \times 10^8}{9.3 \times 10^6 \times 1.60 \times 10^{-19}}$$

$$= 1.3 \times 10^{-13}\,\text{m}$$

64 Green light has a wavelength of 550 nm. At what wavelength do photons have five times as much momentum?

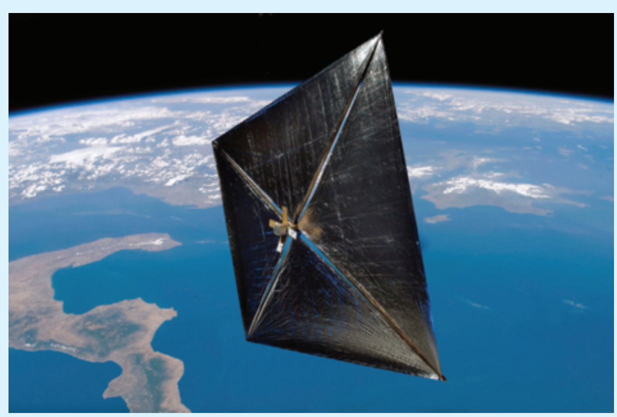
65 What is the momentum of a gamma ray with an energy of 100 keV?

66 When an electron and a positron annihilate, the most common outcome is the production of two gamma rays. If the electron–positron pair were initially stationary then what is the momentum of each of the gamma rays produced?

67 A pion is measured to have a total energy of 550 MeV before it decays into two gamma rays.

a What are the energies of the two gamma ray photons produced?

b What are the wavelengths of the two photons produced?



■ **Figure 13.39** A solar sail is an enormous reflector that uses the solar wind as a propulsion mechanism. Although the total force exerted by the solar radiation is small, over a long period of time it is sufficient to propel space probes to high velocities

68 One of the propulsion systems proposed for space travel is the use of a solar sail (Figure 13.39). At the orbital distance of the Earth the photons arriving from the sun have a total intensity of $1.40 \times 10^3\,\text{W m}^{-2}$. Calculate the momentum of the solar radiation that passes through an area of $1.00\,\text{m}^2$ every second at this distance from the Sun.

69 If a solar sail propulsion system reflects the photons, has a sail area of $15.2\,\text{km}^2$ and has a total mass of $2.00 \times 10^4\,\text{kg}$, what is the acceleration of the space probe?

70 Explain why an effective solar sail must be both highly reflective and have very low density.

ToK Link

In what ways do laws in the natural sciences differ from laws in economics?

The laws of science that are been presented in this chapter have been developed as a result of the failure of previous models to explain experimental results. The laws of relativity were able to explain the reason for the breakdown of the previous models but, importantly, they also made precise predictions that could then be tested experimentally. Einstein's fame is due to the great success of these predictions in matching the later experimental results. Today's theoretical physicists continue to extend the predictions of the laws of relativity while experimental physicists continue to test these predictions. The laws of relativity therefore have an exactness that is missing from the laws of economics.

13.5 (A5: Additional Higher) General relativity –

general relativity is a framework of ideas, applied to bring together fundamental concepts of mass, space and time in order to describe the fate of the universe

Although *special relativity* was an outstanding achievement, Einstein was deeply frustrated that it could only be applied to the special situation of inertial observers. After publishing his work on special relativity he spent much of the next 11 years trying to generalise the theory of relativity and finally published his theory of **general relativity** in 1916. It is based on some remarkably simple ideas although the mathematics behind it is quite complex.

The equivalence principle

Einstein was intrigued that the inertial mass of an object ($m = F/a$) and the gravitational mass of an object ($m = W/g$) are always identical. This means that, if an observer is placed in a closed box, they have no way of distinguishing whether they are in a rocket that is accelerating in the middle of empty space or experiencing a gravitational field. Einstein decided that this was a fundamental property of the universe – physically there was absolutely no measurable difference between a gravitational field strength and an acceleration. We call this the **relativistic equivalence principle**.

Equivalence principle: no experiment can be done to prove that an observer is in an accelerating reference frame or in a gravitational field.

This is illustrated in Figure 13.40. The reason this is so important is because it means that the physics that occurs in one situation must be the same as the physics that occurs in the other.

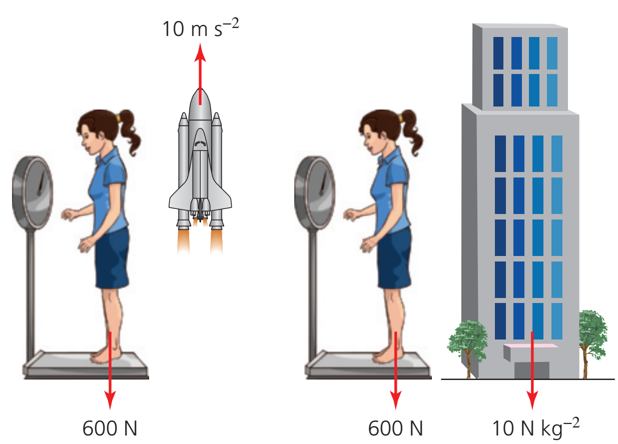


Figure 13.40 The equivalence principle means that, fundamentally, a physicist in a closed box has no way of proving that they are in a room on Earth within a gravitational field of $10\,\text{N kg}^{-1}$ (as shown on the right) or in deep space accelerating at $10\,\text{m s}^{-2}$ (as shown on the left) – or an equivalent combination of the two

The bending of light

Let us consider what happens to a beam of light that is shone by a laser mounted inside the wall of the closed box. In each case the light beam is fired horizontally according to the observer outside the box. The box is drawn after three equal time intervals with each position of the photon as it passes across the box marked by a red dot, acting as an event in spacetime. To show the shape of the photon's path these points are joined up. This is illustrated by Figures 13.41 to 13.46:

1 Situation 1: stationary box – both the internal observer in the box's frame of reference and external observer will agree on the position of the box and the path that the light beam has taken (Figure 13.41):

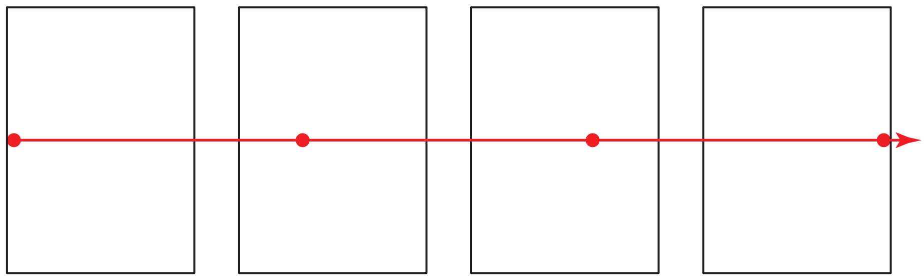


Figure 13.41

2 Situation 2: box moving upwards with constant velocity, v, as seen by the external observer who would see the light beam travelling horizontally and the box travelling at a constant velocity across its path (Figure 13.42):

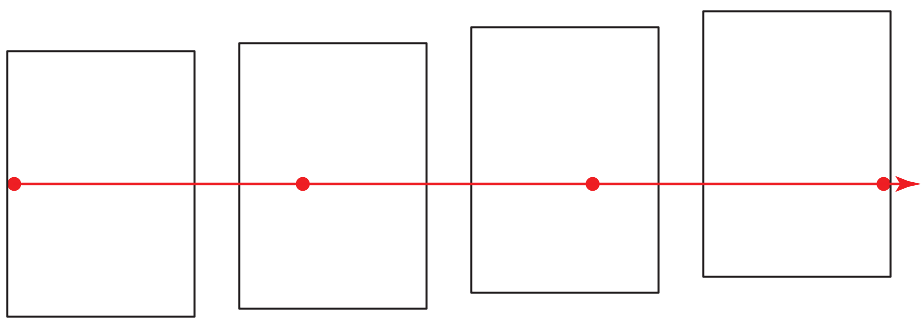


Figure 13.42

However, an observer in the rest frame of the box would see the light beam angled downwards (Figure 13.43):

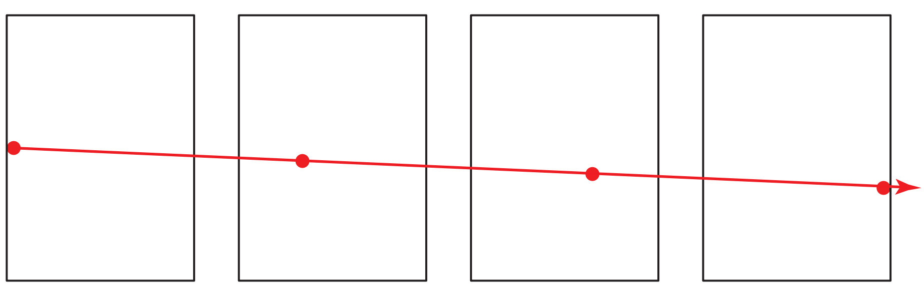
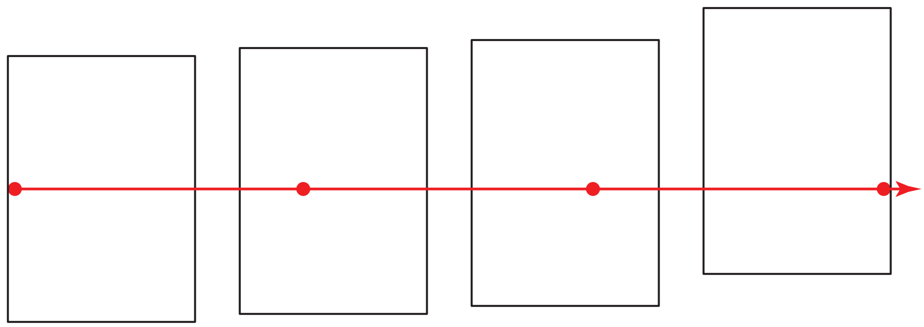


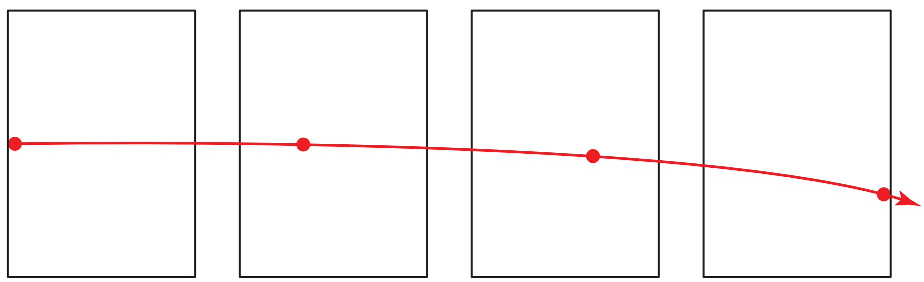
Figure 13.43

3 Situation 3: the box is experiencing a very large upwards acceleration. The external observer would still see the light beam travelling horizontally and the box accelerating across the path of the light beam (Figure 13.44):



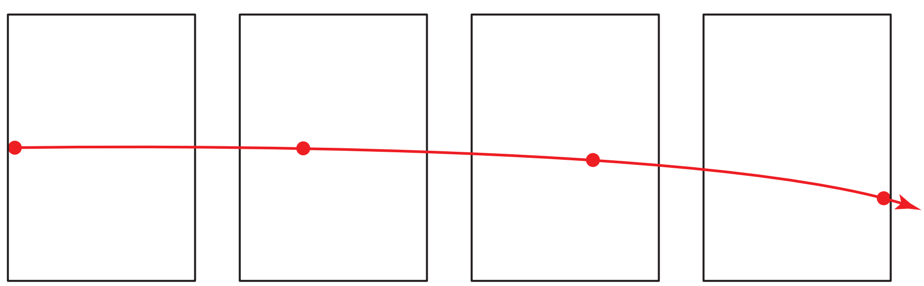
■ **Figure 13.44**

4 Situation 4: the box is experiencing a very large upwards acceleration. This drawn as viewed according to the observer in the rest reference frame of the box. The observer then sees the light beam being bent downwards (Figure 13.45):



■ **Figure 13.45**

5 Situation 5: the box is stationary but in a very strong gravitational field. The equivalence principle can now be applied because the position of the events that make up the path of the light ray must be the same as in situations 3 and 4 above relative to the box (Figure 13.46):

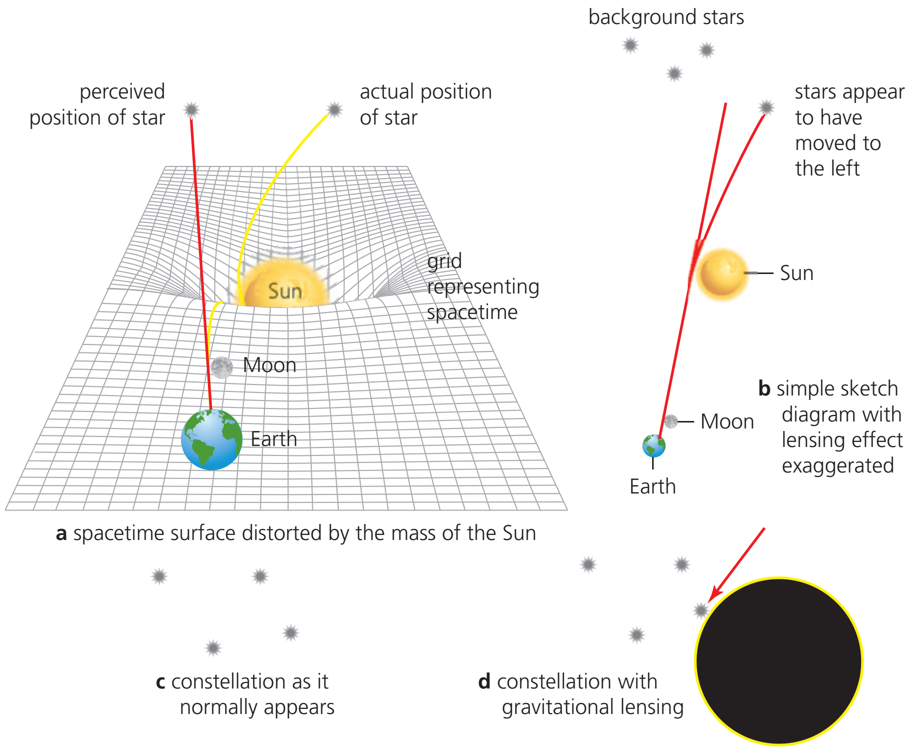


■ **Figure 13.46**

The remarkable conclusion is that gravitational fields must bend the path of a light beam.

■ Evidence for the gravitational bending of light

One of Einstein's early predictions of general relativity was that light must be bent around very massive objects. This is surprisingly difficult to test for two reasons: firstly, the angle through which light is bent is extremely small; secondly, the Sun is so bright that we can only look very close to the edge of Sun during a solar eclipse, when the Moon is directly between the Earth and the Sun. The experiment was first achieved by a team led by Arthur Eddington, using the solar eclipse of 1919 to measure the position of stars close to the edge of the eclipse. This is shown in Figure 13.47.



■ **Figure 13.47** Diagrams showing how gravitational lensing due to the Sun occurs: a) shows a spacetime surface with the yellow line taking the shortest path through spacetime. This is discussed in more detail later; b) shows a simple sketch diagram of the gravitational lensing effect such as you might be asked to produce in an exam question; c) and d) show how a constellation would appear normally and how the constellation would appear when very close to the Sun. The observed position of the shifted star is so close to the edge of the Sun that the effect is only observable from Earth during a solar eclipse, when the moon masks the solar disc. The star marked with the arrow is the one that has shifted position

Gravitational lensing

A *gravitational lens* is formed when the light from a very distant, bright source is 'bent' around a massive object (such as a cluster of galaxies or black hole) between the source object and the observer. The process is known as **gravitational lensing** and is one of the consequences of gravity bending the path of light rays. Einstein discussed the effect in his 1936 paper and it was first experimentally observed in 1979. Figure 13.48 shows an example of what gravitational lensing actually looks like.

■ **Figure 13.48** Gravitational lensing: a very large galaxy (just right of centre) is bending the pathways of light from a source that is behind it from our viewpoint, so that the light appears as a ring known as an Einstein ring

The reason it is called a gravitational lens is because the effect looks similar to the refraction of light through a convex lens, as shown in Figure 13.49.

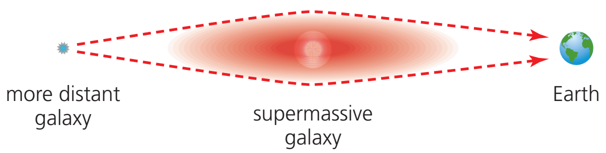


Figure 13.49 Gravitational lensing: the light from a bright distant galaxy or super-bright single star, such as a pulsar, is bent inwards by the strong gravitational field of a supermassive galaxy that lies between a distant galaxy and Earth. The paths of the rays are similar to those produced by a convex lens. In reality the distortion in gravitational lensing often produces either multiple images or a smeared Einstein Ring of the more distant galaxy is produced

Gravitational redshift and the Pound–Rebka–Snider experiment

The bending of light by gravity might suggest that the gravitational field is causing the light ray to change speed, but this would conflict with the first postulate of special relativity (the speed of light in a vacuum is always *c*). Instead, Einstein realised that *time* is being distorted. In an accelerating reference frame, we know that the faster the box is travelling the slower time passes for it, because time passes slower for faster-moving observers than for slower-moving observers. The equivalence principle tells us that the same must be true for a gravitational field – the stronger the gravitational field the slower time must pass. Therefore gravitational fields also cause time dilation – clocks in a strong gravitational field tick slower than clocks in a weak gravitational field.

This result also affects light rays as they pass into an increasing gravitational field, causing them to appear blue-shifted. Conversely, a photon that travels out from a gravitational field will be redshifted, as shown in Figure 13.50 (the Doppler effect and redshifts were discussed in Chapter 9).

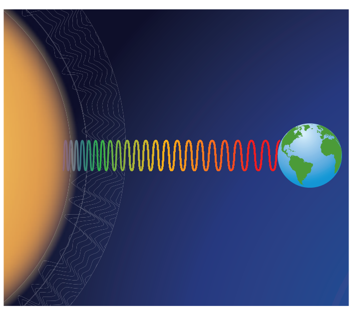


Figure 13.50 The frequency of light is blueshifted as the photon travels into a gravitational field, such as the gravitational field around the Sun and, conversely, the photon is redshifted as it travels out through the *gravitational field*; it is as if the photon is gaining or losing gravitational potential energy, causing its frequency to change

The following equation allows us to calculate the change in frequency, Δf, of the photon:

$$\frac{\Delta f}{f} = \frac{g\Delta h}{c^2}$$

In this f is the original frequency, g is the gravitational field strength, Δh is the change in height and c is the speed of light in a vacuum. The equation assumes that the gravitational field strength is uniform and therefore constant.

This equation is given in the *Physics data booklet*.

Worked example

21 What is the change of frequency of a red light photon, $\lambda = 650\,\text{nm}$ as it travels from the bottom of the Eiffel Tower to the top, a distance of 324 m?

$$f = \frac{c}{\lambda} = \frac{3 \times 10^8}{650 \times 10^{-9}} = 4.62 \times 10^{14}\,\text{Hz}$$

$$\frac{\Delta f}{f} = \frac{g\Delta h}{c^2}$$

$$\Delta f = \frac{9.81 \times 324 \times 4.62 \times 10^{14}}{(3.00 \times 10^8)^2}$$

$$= 16.2\,\text{Hz}$$

As can be seen, the change in the frequency compared with the original frequency is tiny, making this a particularly challenging prediction of general relativity to test.

The Pound–Rebka–Snider experiment

The key experimental test of gravitational redshifting happened in an experiment at Harvard University in 1959 and is named after the three physicists who designed and refined the original experiment.

According to quantum physics only very specific photon energies can be emitted and absorbed by electrons as they jump between allowed energy levels. The same is true for the much higher energies for the energy states of unstable nuclei that are involved in the emission and absorption of gamma rays – when a nucleus changes from a high-energy state to a low-energy state it emits gamma rays of a specific energy (and frequency). Conversely, when a gamma ray of exactly the right energy interacts with a nucleus it can be absorbed, causing the nucleus to jump from a lower to a higher from a lower energy state. Gamma rays that do not have exactly the right energy will pass through the sample without being absorbed. This provided the evidence for nuclear energy levels, which was discussed in Chapter 12.

Pound and Rebka placed a sample of iron-57 in the basement of a laboratory and a second sample of iron-57 in a room on the top floor. The two samples needed to be as far apart in the Earth's gravitational field as is conveniently possible to maximise the redshift effect. By focusing the gamma rays emitted by the lower sample into a narrow beam they were able to show that the sample of iron-57 on the top floor could no longer absorb the incident gamma rays. The gravitational redshifting of the gamma rays was sufficient that the rays emitted by one sample could not be absorbed by the other. However, to test the predictions of general relativity they needed to test how large the gravitational redshift was. They did this by initially placing the basement sample on a loudspeaker, which they oscillated up and down. This resulted in a Doppler blueshift when the sample moved upwards and a Doppler redshift when the speaker moved downwards.

At a specific point in the phase of the speaker oscillation, the sample on the top floor was able to absorb the gamma rays from the sample in the basement. This occurs when the gravitational redshift exactly cancels the Doppler blueshift. The result is measured as a drop in the intensity of the gamma radiation passing through the top sample, showing that gamma radiation had been absorbed. By working out the exact point in the oscillation of the speaker that allows absorption to occur, Pound and Rebka were able to calculate the velocity of the speaker and, hence, the Doppler blueshift. They repeated the experiment in reverse to test that the gamma photons that travelled from the top floor to the basement were also gravitationally blueshifted when travelling downwards. Their results confirmed that the gravitational redshift and blueshift matched that predicted by general relativity. Pound and Snider went on to repeat the experiment with greater accuracy; hence all three are recognised in the name of the experiment. This is shown in Figures 13.51 and 13.52.

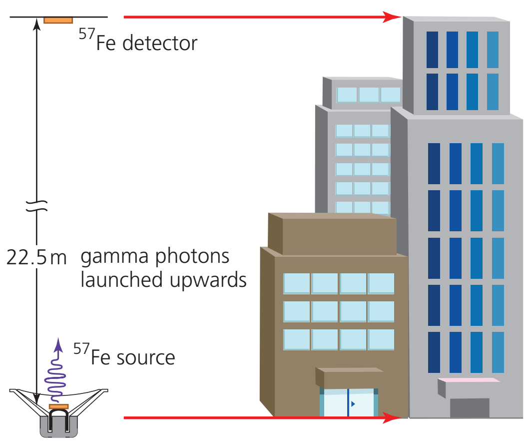


Figure 13.51 The Pound–Rebka–Snider experiment

Figure 13.52 Rebka at work in the basement of the Jefferson laboratory at Harvard; the polythene tube contains helium to focus the gamma photons into a narrow beam

71 Images from the Hubble Space Telescope show what appears to be three very hot, bright indentical stars around a dark region of space. The three stars all appear to be the same distance away while the dark region appears to be filled by a much closer dust cloud emitting light in the infrared spectrum. Explain why there appears to be three images of the same star.

72 In the Pound–Rebka–Snider experiment, the Jefferson Laboratory in which the experiment was carried out had a height of 22.6 m between the two samples. The gamma rays produced by iron-57 had an energy of 14.0 keV.

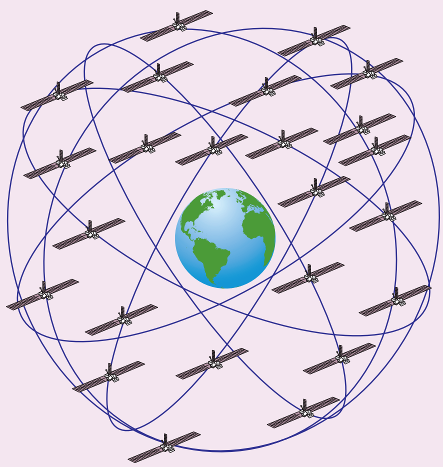
- **a** Calculate the gravitational redshift of the photons as they lost energy while rising up through the gravitational field between the source and the detector samples.
- **b** In order for the photons to be absorbed by the sample of iron-57 on the top floor they must have had exactly the right energy and so must have been blueshifted by the oscillating speaker in the basement. Use the Doppler shift equation for light to calculate the velocity of the speaker.

73 The darker lines on the absorption spectra of the Sun do not exactly correspond to the emission spectra lines of gases on Earth but appear shifted slightly. Explain what is causing this effect.

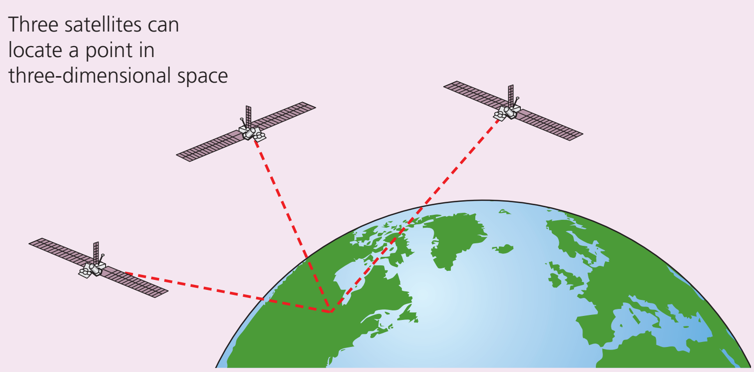
Utilizations

GPS time correction

In 1956 a test was proposed for relativity that involved placing very accurate clocks in orbiting satellites and measuring the increase in the rate of the clocks due to the combination of special and general relativistic effects. Forty years later the Global Positioning System (GPS), launched by the American government, had to account for this very effect (Figures 13.53 and 13.54). Each of the GPS satellites carries an atomic clock that works with a precision of ±1 nanosecond. At the speed of light, *c*, this results in a precision in position of ±0.3 m.

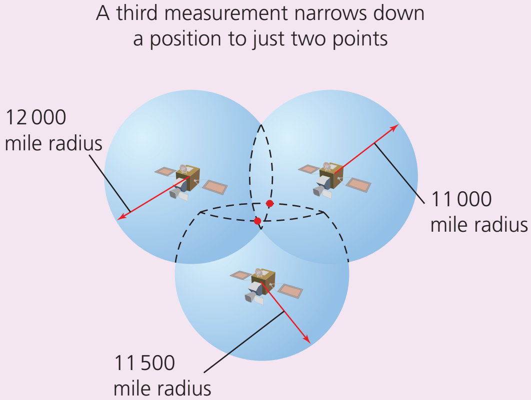


■ **Figure 13.53** There are now more than 24 GPS satellites in orbit around the Earth



■ **Figure 13.54** Fixes from each of three satellites are sufficient to provide a position fix, but a fourth satellite is needed to account for inaccuracies in the GPS receiver's clock and to ensure that common electronic devices, such as mobile phones and cameras, can easily provide accurate positioning

As each GPS satellite orbits it continually sends out a signal providing information about its position and orbital path, with a precise time signal stating when the location information was sent. The signal travels out in a sphere around the satellite centred on the position of the satellite. The receiver can work out the radius of the sphere by calculating the time the signal has been travelling for at the speed of light. If three GPS satellites are all compared then these spheres should all intersect at the location of the GPS receiver allowing the GPS receiver to calculate an exact position, as shown in Figure 13.55. However, GPS receivers do not require atomic clocks to calculate the time correctly; instead they use a fourth satellite to measure the time discrepancy and, from this, accurate measurements of the time between the signals being sent and received are calculated.



■ **Figure 13.55** Assuming the receiver has an accurate time measurement, the position and time measurements from three GPS satellites provide enough information to calculate the receiver's position to one of two locations, shown as the red dots where the spherical surfaces intersect. In practice, one of the two locations can be discarded because it is so far from the Earth's surface. Comparison with a fourth satellite provides a time correction so that the receiver does not need to be linked to an atomic clock

The high speed of GPS satellites as they orbit around the Earth – approximately 14 000 km h^{-1} – means that time is dilated due to special relativity, so that the atomic clocks run slow by 7000 nanoseconds per day. However, the satellites are also in a weaker gravitational field due to their altitude above the Earth's surface. General relativity implies that their clocks should therefore run faster. At the GPS satellite orbital height they are predicted to run 45 000 nanoseconds faster per day. The combined effect is that the satellites should run fast by around 38 microseconds each day which, at the speed of light, would create a daily inaccuracy of 11.4 km in the GPS positioning.

To get around this problem the incredibly precise clocks in the GPS satellites are deliberately set to tick 38 microseconds too slow each day. Not only do GPS satellites provide further evidence supporting relativity but, without relativistic corrections, none of the satellite navigation systems that many of us routinely use would work.

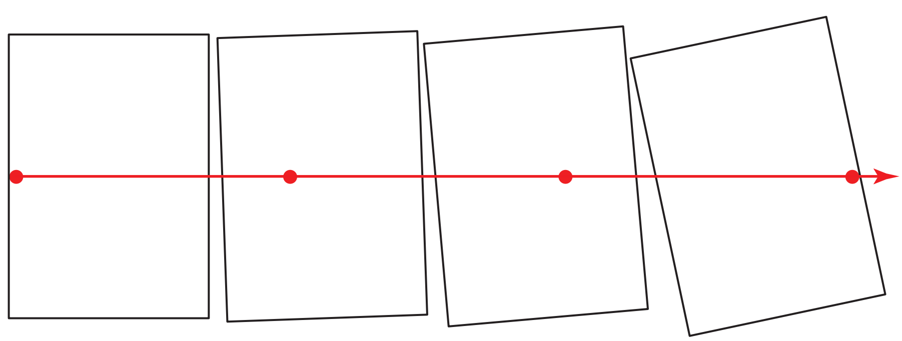
74 If the clocks on the GPS satellites ran fast by only 1 ns per day what would be the cumulative error in the distance to the satellite after:
- **a** 1 day?
- **b** 1 week?
- **c** 5 years?

75 The measurement of position depends on the assumptions that the clocks are correct and that the positions of the satellites are correctly broadcast. Suggest other sources of error in GPS measurements.

76 Who was responsible for carrying out the experiment in 1971 that showed that clocks would run faster at higher altitudes?

Curvature of spacetime

In describing special relativity Einstein realised that an observer's relative velocity caused a change in how space and time behaved. It is only a small step to realise that the distortion of spacetime increases as the observer accelerates, causing spacetime to become curved. The equivalence principle then tells us that gravitational fields also curve spacetime. In other words, high densities of mass (and therefore energy) must stretch and distort spacetime. Einstein also thought that light must take the shortest, and therefore straightest, path available across spacetime. This is shown in Figure 13.56.

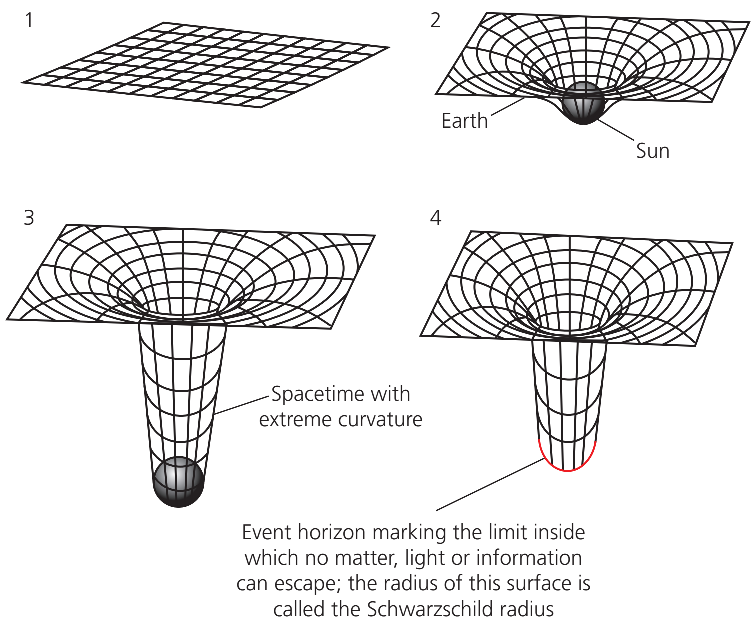


■ **Figure 13.56** It is spacetime that is being distorted, producing a curving and stretching of space and time. Although the light beam appears to follow a curve through the three dimensions of space, it is still travelling as the straightest line through spacetime. Spacetime itself is being stretched and curved, and will cause the box to appear to rotate and be distorted

The first law of Newtonian physics is that an object will travel along a curved path only as a result of an external unbalanced force. Matter will travel in a straight line until an unbalanced force acts on it. However, when spacetime is curved, matter will follow this curve and experience an unbalanced force as a result.

Einstein hypothesised that what we experience as the force of gravity is the same as spacetime causing an object to travel in a curve. One way of thinking about this is to suggest that energy and mass have the effect of defining how spacetime must curve, while the curvature of spacetime has the effect of determining how mass and energy will move.

This resolved one of Einstein's biggest problems with special relativity, because he realised that gravity appears to instantaneously exert a changing force on an orbiting object at a distance, apparently transmitting information about the direction of the force on the orbiting object faster than the speed of light. Instead the curvature of spacetime around a large gravitational body causes orbiting objects to follow a curved orbital path. This curvature of spacetime is shown diagrammatically in Figure 13.57.



■ **Figure 13.57** A commonly used analogy is to think of a two-dimensional representation of spacetime as a rubber sheet that is stretched in all directions in a horizontal plane. Massive objects are represented by placing heavy balls on the sheet with both mass and density being represented by more massive and denser balls respectively. However, the rubber sheet can be torn if the density reaches a critical limit, at which spacetime becomes so stretched that light is unable to escape – the resulting object is called a *black hole*

■ Event horizons

If mass (or energy) is concentrated into a superdense object then the gravitational field strength close to the object's surface also increases. This means that the escape speed also increases. However, special relativity determines an upper limit for speed. The escape speed limit discussed in Figure 13.57 occurs when the escape speed (see Chapter 10) exactly equals the speed of light in a vacuum. This means that light is unable to escape from within this region.

We call the surface that joins all the points at which the escape speed = c the *event horizon*. This is a theoretical surface that separates a superdense object such as a black hole from the rest of the universe. The event horizon marks the surface within which light, and therefore all forms of matter and also information, is unable to escape through to the rest of the universe.

Schwarzschild black holes

Schwarzschild black holes are simplest form of black holes and are neither rotating nor have an electrical charge. The critical limit discussed in Figure 13.57 occurs when the escape velocity (Chapter 10) exactly equals the speed of light in a vacuum. This means that light is unable to escape from within this region. The radius that this occurs at for a simple *non-rotating black hole with no electrical charge* is called the Schwarzschild radius, R_S, and is given by:

$$R_S = \frac{2GM}{c^2}$$

Where G is the universal gravitational constant, M is the mass of the black hole and c is the speed of light in a vacuum. This equation is given in the *Physics data booklet*.

The Schwarzschild radius defines a spherical surface in space. Inside this boundary no energy or information can escape into the rest of the universe while outside this spherical surface it can. The spherical surface is therefore the black hole's **event horizon**.

Worked example

22 What would be the radius of the event horizon of an object with the mass of the Sun if it were sufficiently dense to produce a black hole? The Sun has a mass of 1.99×10^{30} kg.

$$R_S = \frac{2GM}{c^2}$$

$$= \frac{2 \times 6.67 \times 10^{-11} \times 1.99 \times 10^{30}}{(3 \times 10^8)^2}$$

$$= 2.95 \times 10^3 \text{ m}$$

In reality, the only force understood to produce dense enough mass to create a stellar black hole is gravity. The unbalanced inward force is only sufficient to crush the star into a black hole at the end of its life, and is only formed by extremely large stars (red supergiants)with a final mass of no less than 1.5 to 3.0 solar masses and initial masses of 15 to 20 solar masses.

Nature of Science

Creative and critical thinking

Einstein's great achievement – the general theory of relativity – connecting the geometry of spacetime (through its curvature) to the mass and energy content of spacetime is based on intuition, creative thinking and imagination.

For years it was thought that nothing could escape from a black hole and this is true but only for classical black holes. When quantum theory is taken into account a black hole radiates like a black body. This unexpected result revealed other equally unexpected connections between black holes and thermodynamics.

In 1974 Stephen Hawking, now Director of Research at the Centre for Theoretical Cosmology at the University of Cambridge, was trying to link quantum physics with general relativity. As part of his research he was investigating the quantum field effects close to the event horizon. Mathematically this is very complex but, put very simply, quantum physics predicts the existence of virtual particles caused by quantum fluctuations in the vacuum energy. Due to the huge gravitational field these particles can become real particles. Theoretically it is possible for one of these particles to fall through the event horizon while the other particle escapes, and because the energy to produce the particle has come from the gravitational field, the escaping particle carries away some of the black hole's energy (or mass).

The result is that black holes should appear to glow as if they were a hot black-body. However, they glow with a temperature inversely proportional to their mass. Small black holes therefore appear hotter than large black holes. Indeed, a black hole with the mass of our Moon (10^{27} kg) would glow at around 2.7 K, the temperature of the microwave background radiation, and so any more massive than this and they will still absorb more radiation than they emit.

Intriguingly, because the smaller a black hole gets the hotter it gets, the rate of evaporation increases exponentially as the black hole shrinks. This result means that miniature black holes could appear and quickly evaporate through Hawking radiation. Indeed, the theory suggests that the LHC at CERN might have sufficient energy to create such miniature black holes and the ability to watch them evaporate.

Time dilation near a black hole

Both space and time are stretched by the mass of a black hole so, relative to an external observer far from the black hole, time will appear to become increasingly stretched (dilated) as an object approaches the event horizon. We can calculate the distortion in the time between two events according to an observer close to the black hole and an observer far from the black hole. This is given by the equation:

$$\Delta t = \frac{\Delta t_0}{\sqrt{1 - \frac{R_S}{r}}}$$

Where Δt_0 is the time between two events according to an observer close to the black hole, Δt is the time between two events according to an observer far from the black hole, R_S is the Schwarzschild radius of the black hole and r is the distance of the closer observer from the black hole. This equation is given in the *Physics data booklet*.

Worked example

23 A space probe is sent to investigate a black hole with a Schwarzschild radius of 5.0 km and sends a signal with a pulse of period 0.01 s back to Earth. What is the period of the signal when the space probe is 10 km from the centre of the black hole?

$$\Delta t = \frac{\Delta t_0}{\sqrt{1 - \frac{R_S}{r}}}$$

$$= \frac{0.01}{\sqrt{1 - \frac{5000}{10000}}}$$

$$= 0.014\,\text{s}$$

77 In December 2011 astronomers discovered a supermassive black hole in the centre of the Milky Way that has an estimated mass of 21 billion solar masses (4.2×10^{40} kg). Estimate the Schwarzschild radius of this black hole.

78 A pulsar, which is a form of magnetised spinning neutron star, emits a beam of very intense radiation that appears to flash with a regular period, like a lighthouse. Pulsars spin very quickly and can have a period of milliseconds. If a 33 millisecond pulsar is found to orbit around a black hole with a mass of 3.0×10^{35} kg so that it comes close enough so that the period is increased, due to general relativistic effects, to 34 milliseconds, how far is the neutron star from the centre of the black hole?

Applications of general relativity to the universe as a whole

Einstein realised that if spacetime could be curved and distorted then the universe as a whole could also have curvature. The classical model of the universe is what we call Euclidean. The three dimensions of space are orthogonal – at right-angles – completely straight with uniform scales in all directions. This is an example of an 'open universe' because it is infinite in at least one dimension. General relativity allows the dimensions of time and space to be curved so that opposite directions in space, or in time, can eventually curve around and meet up.

The surface of the Earth is a good example of a closed surface – it is so large that on the human scale it is reasonable to think of the Earth as being flat (and therefore either being infinite or having an edge to fall off). Instead the Earth's surface is curved, so it does not have a start or a finish. Curving space and time means that the universe can also become closed, so the problem of what is beyond the edge of the universe disappears.

Utilizations

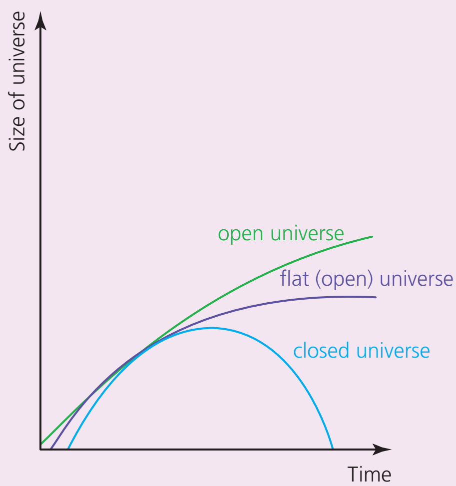
Implications of a relativistic universe: expansion and contraction

Einstein applied the laws of relativity to the universe as a whole by treating the galaxies like molecules in a cosmological gas and applied the statistical rules of thermodynamics to model how the universe behaves.

He realised that his relativistic universe was intrinsically unstable. Galaxies must attract each other causing both space and time to collapse. His solution was to introduce an outward force, which he called the *cosmological constant*, representing a repulsive force in the universe. The cosmological constant gives the vacuum (or each unit of spacetime) a tiny amount of energy. Too much energy and the universe will expand at an increasing rate; too little energy and gravity overcomes the repulsion and the universe collapses.

In 1929 Edwin Hubble announced that the universe was expanding. The invention of larger telescopes fitted with cameras has allowed astronomers to look much deeper into the universe and Hubble had discovered that most of the galaxies that he could detect appear to be redshifted, implying that they are moving away from us. Importantly, the amount of redshifting increases with distance. His conclusion was that the universe must be getting larger.

The implication was that the universe started from a single point of incredibly high temperature and density and has expanded outwards ever since. We call this event the Big Bang and it is now a fundamental aspect of modern cosmology. After the Big Bang event our models of the universe suggest a rapidly expanding cosmos that cools as it expands. The assumption used to be that gravity would slow down the rate of expansion of the universe and, if there was sufficient mass in the universe, it could even cause the universe to ultimately collapse – otherwise it would continue expanding at a slowing rate forever. Our universe is not stable – it is continually evolving – and until recently the possibilities were considered to be as shown by the graph in Figure 13.58.



■ **Figure 13.58** The three simplest solutions to the general relativity field equations for the evolution of the universe with different mass/energies

Einstein swiftly withdrew the cosmological constant from his original field equations and famously said that introducing it was his biggest blunder. Ironically, he might have been completely wrong because in 1998 two research teams independently came to the conclusion that the rate of expansion of the universe was not slowing but was actually increasing (not shown in Figure 13.58). Saul Perlmutter, Brian Schmidt and Adam Riess jointly received the Nobel prize for physics in 2011 for their work on measuring supernovae events. The most distant supernovae were dimmer than expected, suggesting that they were further away than a slowing expansion suggests they should be. The suggestion has been that space must be filled with a dark energy somewhat similar to Einstein's cosmological constant.

1 Explain why it is so difficult to measure whether spacetime is curved or flat on cosmological scales.
2 Open, flat and closed universes have different outcomes. What are the likely outcomes for each type of universe?
3 Explain why the age of a universe also depends on whether the universe is open, flat or closed.
4 In cosmology, *inflation theory* also uses a similar concept to the cosmological constant to explain the large-scale structure of the universe from the quantum fluctuations of the early universe. What is inflation theory?

ToK Link

What other examples are there of initially doubted claims being proven correct later in history?

Although Einstein self-described the cosmological constant as his 'greatest blunder', the 2011 Nobel prize was won by scientists who had proved it to be valid through their studies on dark energy.

Students will no doubt be questioning what is required to prove that an idea is valid. Human understanding has been a succession of new models that are at first doubted and then, when increasing experimental data verifies them, acclaimed. Ideas such as evolution, plate tectonics, the spherical nature of the Earth, the theory of thermodynamics and even climate change have all initially challenged the established views and have had to overcome considerable doubt to become mainstream ideas. In contrast are claims that are appealing or appear to be commonsense and are therefore assumed to be valid without supporting evidence. Suggested areas of worthwhile discussion are educational learning styles, homeopathy and intelligent design.

Summary of knowledge

13.1 The beginnings of relativity

- Reference frames are systems of coordinates for different moving observers.
- Newton's postulates state that time ticks at a constant rate everywhere in the universe so that all observers agree on the time interval between two events.
- 'Simultaneous' means that the events occur with no time interval between them.
- Galilean transformation equations ($x' = x - vt$ and $u' = u - v$) provide a means of transforming the coordinates of one reference frame to another.
- Maxwell's equations explain how electrical charges, magnets, electrical fields and magnetic fields interact.
- Maxwell's equations derive a constant value for the speed of light in a vacuum, which contradicts Newtonian physics.
- The existence of electrical and magnetic forces on charges and currents moving through magnetic and electrical fields depends on the reference frame of the observer, which contradicts Newtonian physics.

13.2 Lorentz transformations

- The definition of an inertial reference frame – a coordinate system that is neither accelerating nor experiencing a gravitational field.
- The two postulates of special relativity – all inertial reference frames are equally valid; the speed of light in a vacuum is a constant for all inertial observers.
- Clocks can be synchronised by a stationary observer standing midway between two stationary clocks using a flash to send a beam of light to each clock.
- Events that are simultaneous for one observer do not have to be simultaneous for another observer.
- Spacetime is the fundamental entwining of space and time so that relative motion affects an observer's measurements and experience of both space and time.
- Invariant quantities, that is quantities that are constant for all observers – the speed of light in a vacuum, spacetime interval, proper time interval, proper length and rest mass.
- The spacetime interval is the separation between events in spacetime.
- The spacetime interval is an invariant quantity so it must be the same for all observers:

$$(\Delta s)^2 = c^2 (\Delta t)^2 - (\Delta x)^2$$

- Rest mass, m_0, is the mass of an object measured by an observer who is stationary relative to the object.
- Proper time interval, Δt_0, is the time measured between two events by an observer who experiences both events occurring in the same place in their coordinate system (reference frame).
- Proper length, L_0, is the length measured between two events by an observer who is not moving relative to the two events.
- The Lorentz factor describes the scaling that occurs when transforming between one reference frame and another:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

- Lorentz transformations describe mathematically how to change coordinates when moving from one inertial reference frame to another:

$$x' = \gamma(x - vt)$$

$$t' = \gamma\left(t - \frac{vx}{c^2}\right)$$

- Lorentz transformations can also be used to describe how the length between two events or the time interval between two events changes depending on the relative motion of the observer:

$$\Delta x' = \gamma(\Delta x - v\Delta t)$$

$$\Delta t' = \gamma\left(\Delta t - \frac{v\Delta x}{c^2}\right)$$

- The lightyear, ly, is a unit of distance equal to the distance travelled by light in a vacuum for one year.
- When transferring from one reference frame to another, relative speeds do not simply add up. Instead, the Lorentz velocity transformation equation must be used:

$$u' = \frac{u - v}{1 - \frac{uv}{c^2}}$$

- Time dilation is the stretching of time as a result of relative motion. The proper time interval is the shortest possible time interval between two events with all other observers measuring a longer time interval calculated from the equation:

$$\Delta t' = \gamma \Delta t_0$$

- Length contraction is the shortening of the distance between two events due to relative motion. The proper length is the longest possible straight length between two events with all other observers measuring a contracted length calculated from the equation:

$$L = \frac{L_0}{\gamma}$$

- The muon-decay experiment provided one of the tests of special relativity. Muons are formed in the atmosphere as a result of collisions with cosmic rays. They have a very short half-life so, according to Newtonian physics, very few should reach the Earth's surface. The numbers that do arrive correspond with the effects of time dilation and length contraction due to the muon's high speed.

13.3 Spacetime diagrams

- Spacetime diagrams represent the dimension of space, x, along the horizontal axis while the vertical axis represents either time or, more commonly, the speed of light in a vacuum multiplied by time, ct.
- Events are represented as points on a spacetime diagram.
- A line of events represents the path taken through spacetime by an object – this is called the object's worldline.
- For photons, the worldline is usually drawn at 45° and is sometimes called a lightline.
- The gradient of a worldline is given by c/v.
- The angle between a worldline and the vertical axis is given by:

$$\theta = \tan^{-1}\frac{v}{c}$$

- A second reference frame can be represented on the same spacetime diagram, but the new ct' axis is along the worldline of a particle that is at rest in this reference frame.
- The new x' axis mirrors the new ct' axis along the lightline (the worldline of a photon).

- This produces a skewed pair of axes. Coordinates in this reference frame can be read from a skewed grid that is drawn parallel to the x' and ct' axes.
- Although the background spacetime is the same for both reference frames, the scales on the original and the scaled axes are different.
- Events that are simultaneous for one observer do not have to be simultaneous for observers in different reference frames. Events are simultaneous if they occur along the same grid line parallel with the x or x' axis.
- Time dilation and length contraction can be demonstrated on spacetime diagrams.
- The twin paradox was proposed as an argument against special relativity. It suggests that when two twins move rapidly apart from each other, each twin thinks it is stationary and that the other is moving, and so the time dilation effects appear to be indentical for each twin. The travelling twin then returns and the situation once again appears symmetrical.
- The paradox is resolvable because the situation is *not* symmetrical. The travelling twin is not in a single inertial reference frame but changes from one inertial frame to another.

13.4 Relative mechanics

- Rest mass and electrical charge are invariant quantities. Total mass can be calculated from the rest mass using the equation:

$$m = \gamma m_0$$

- Energy and mass are different aspects of the same thing. As an object gains energy it also gains mass. The amount of energy possessed by a stationary object is called its rest energy, which is calculated using the equation:

$$E_0 = m_0 c^2$$

- Rest energy is an invariant quantity.
- The total energy of an object is a combination of its rest energy and its kinetic energy. It is calculated using the equation:

$$E = \gamma m_0 c^2$$

- The total energy of a particle tends to infinity as the particle approaches the speed of light. Consequently it is impossible for a particle to reach or exceed the speed of light.
- On a graph of total energy against velocity the y-intercept is the rest energy of the particle. A similar graph can be drawn comparing mass against velocity.
- Kinetic energy can no longer be calculated using the equation $E_k = \frac{1}{2}mv^2$; instead relativistic kinetic energies are calculated by subtracting the rest energy from the total energy:

$$E_k = (\gamma - 1)m_0 c^2$$

- At relativistic speeds the equation $p = mv$ is no longer valid; instead the relativistic version of the equation must be used:

$$p = \gamma m_0 v$$

- A particle's total energy can also be calculated using the equation:

$$E^2 = p^2c^2 + m_0{}^2c^4$$

- When making relativistic calculations it is much simpler to use units written in terms of c. For energy these are still eV, keV, MeV and GeV. For mass these become: eV c^{-2}, keV c^{-2}, MeV c^{-2} and GeV c^{-2}. For momentum the units are: eV c^{-1}, keV c^{-1}, MeV c^{-1} and GeV c^{-1}.
- When particles are accelerated all the energy that they gain is regarded as kinetic energy. Therefore:

$$qV = \Delta E_k$$

- Photons have no rest mass but they do have momentum and energy. The links between a photon's momentum, energy and frequency are given by the equation:

$$p = \frac{E}{c} = \frac{hf}{c} = \frac{h}{\lambda}$$

13.5 General relativity

- The equivalence principle states that the physics is identical whether an observer is in an accelerating reference frame or in a gravitational field. This is useful because it means that what can be deduced about one situation must also be true in the other.
- An observer in an accelerating reference frame will see a light beam curving. Consequently the equivalence principle means that light must also bend in a gravitational field.
- The evidence for the gravitational bending of light and the gravitational distortion of spacetime is provided by gravitational lensing of distant stars by intervening massive galaxies and by the small shift in the position of stars when viewed close to the Sun during a solar eclipse.
- Gravitational redshifting occurs to photons as they escape from a gravitational field. Gravitational blueshifting occurs when photons fall deeper into a gravitational field.
- The amount of redshifting can be calculated from the following equation (g is still used as though the field is uniform):

$$\frac{\Delta f}{f} = \frac{g\Delta h}{c^2}$$

- The Pound–Rebka–Snider experiment used gamma rays produced by iron nuclei to measure the gravitational redshift. The gamma rays could only be absorbed by a source at a higher point when the gravitational redshift was cancelled by a Doppler blueshift caused by oscillating the surface.
- The event horizon is the surface that joins the points where the escape speed matches the speed of light in a vacuum. The event horizon therefore marks the edge of the black hole. Inside light, information and matter are unable to escape into the rest of the universe.
- Mass and energy stretch spacetime. When spacetime is stretched beyond a certain point light can no longer escape – the result is a black hole. For simple black holes that are neither electrically charged nor spinning, the radius of the event horizon is called the Schwarzschild radius and is given by the equation:

$$R_S = \frac{2GM}{c^2}$$

- Gravitational fields also result in gravitational time dilation. The amount of time dilation close to a black hole can be calculated using the equation:

$$\Delta t = \frac{\Delta t_0}{\sqrt{1 - R_S/r}}$$

- General relativity also predicts that the entire fabric of the universe may be distorted so that one or more dimensions are closed or circular.
- The universe is unstable and cannot remain in a static state. The basic model of the cosmos suggests that gravity should cause the universe to slow down in its expansion and. If the universe has sufficient mass, it could collapse back on itself.
- However, recent supernovae data suggest that the expansion rate is increasing.

Examination questions – a selection

Paper 3 IB questions and IB style questions

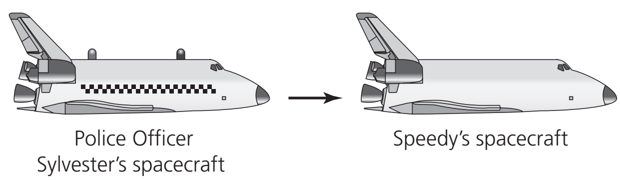
Q1 This question is about the forces acting on moving charges. Two protons are virtually stationary in a laboratory and are placed 10^{-10} m apart.

a Draw a diagram of the two protons, labelling the principal forces acting on each proton. (2)

b The protons are now accelerated perpendicular to the line that separates them so that they travel through a vacuum with initially parallel velocities. Draw a second diagram to show the forces acting on the protons according to an observer in the laboratory reference frame. (3)

c Maxwell's equations predict that the speed of light through a vacuum should be constant. Explain why this contradicts the laws of Newtonian physics. (3)

Q2 This question is about relativistic kinematics. Speedy is in a spacecraft being chased by Police Officer Sylvester.



In Officer Sylvester's frame of reference, Speedy is moving directly towards Officer Sylvester at 0.25*c*.

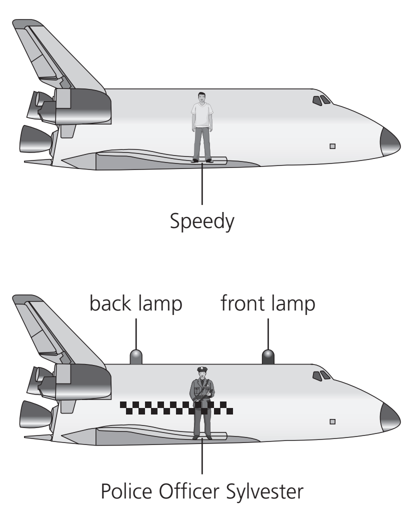
a Describe what is meant by a frame of reference. (2)

b Officer Sylvester switches on the blue flashing lamps on his police spacecraft.

i Assuming that a Galilean transformation applies to this situation, calculate the value of the speed of the light that Speedy would measure using the flashing lamps. (1)

ii Speedy measures the speed of the light emitted by the flashing lamps. Deduce, using the relativistic addition of velocities, that Speedy will obtain a value for the speed of light equal to *c*. (3)

c At a later time the police spacecraft is alongside Speedy's spacecraft. The police spacecraft is overtaking Speedy's spacecraft at a constant velocity.
Officer Sylvester is at a point midway between the flashing lamps, both of which he can see. At the instant when Officer Sylvester and Speedy are opposite each other, Speedy observes that the blue lamps flash simultaneously.
State and explain which lamp is observed to flash first by Officer Sylvester. (4)



d The police spacecraft is travelling at a constant speed of 0.5c relative to Speedy's frame of reference. The light from a flash of one of the lamps travels across the police spacecraft and is reflected back to the light source. Officer Sylvester measures the time taken for the light to return to the source as 1.2×10^{-8} s.

i Outline why the time interval measured by Officer Sylvester is a proper time interval. (1)

ii Determine, as observed by Speedy, the time taken for the light to travel across the police spacecraft and back to its source. (3)

Q3 This question is about special relativity, simultaneity and length contraction.

a One of the two postulates of special relativity may be stated as: 'The laws of physics are the same for all observers in inertial reference frames.' State:

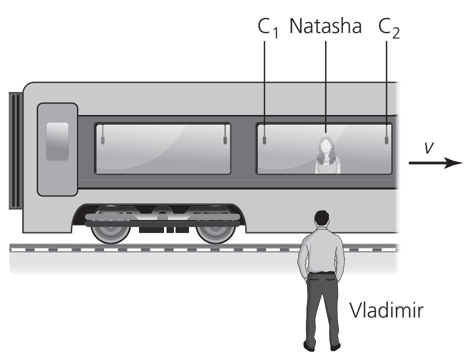
i what is meant by an inertial frame of reference (1)

ii the other postulate of special relativity. (1)

b In a thought experiment to illustrate the concept of simultaneity, Vladimir is standing on the ground close to a straight, level railway track. Natasha is in a railway carriage that is travelling along the railway track with constant speed, v, in the direction shown.

Natasha is sitting on a chair that is equidistant from each end of the carriage. At either end of the carriage are two clocks, C_1 and C_2. Next to Natasha is a switch that, when operated, sends a signal to each clock. The clocks register the time of arrival of the signals. At the instant that Natasha and Vladimir are opposite each other, Natasha operates the switch. According to Natasha, C_1 and C_2 register the same time of arrival of each signal.

Explain, according to Vladimir, whether or not C_1 and C_2 register the same time of arrival for each signal. (4)



c The speed, v, of the carriage is 0.70c. Vladimir measures the length of the table at which Natasha is sitting to be 1.0 m.

i Calculate the length of the table as measured by Natasha. (3)

ii Explain which observer measures the proper length of the table. (1)

d According to Vladimir, a clock at rest in the railway carriage will appear to run slower than a clock at rest beside him. However, according to Natasha, Vladimir's clock will run slower than a clock at rest beside her.

i Outline how this time dilation phenomenon leads to the 'twin paradox', in which one of the twins embarks on a return journey to a distant star at a speed close to that of light while the other twin remains on Earth. (3)

ii State the reason behind the resolution of the paradox. (1)

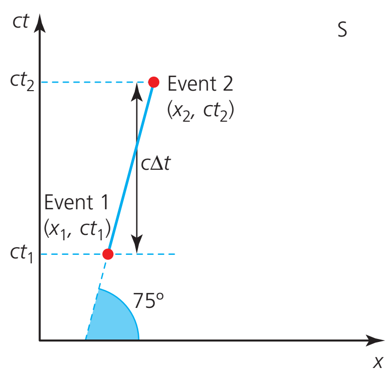
e Evidence for time dilation comes from the decay of muons. A pulse of muons produced by cosmic radiation in the upper atmosphere of the Earth travels to Earth with a speed of 0.96c as measured by an observer at rest on the surface of the Earth. The half-life of the muons, as measured in the frame of reference in which the muons are at rest, is 3.1×10^{-6} s.

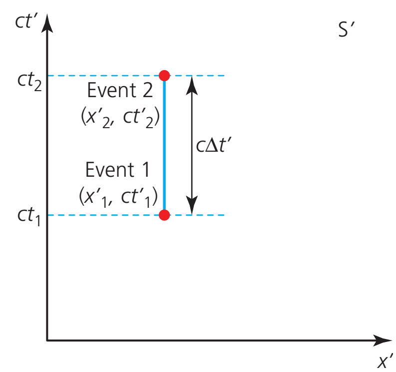
i Determine, for the muons, the distance that Earth will have travelled towards them after half of the muons in the pulse have decayed. (1)

ii Calculate, for the Earth observer, the distance that the muon pulse will have travelled towards Earth after half of the muons in the pulse have decayed. (2)

f Suggest how your answers to **e i** and **e ii** provide evidence that supports the theory of special relativity. (3)

Q4 This question is about spacetime diagrams. Look at the two spacetime diagrams below. The solid blue lines represent the worldlines of a muon that is formed at Event 1 and decays into two photons at Event 2. The two spacetime diagrams have been drawn so that they have identical scales although they start at different times.





a Which reference frame is the rest reference frame of the muon? (1)

b Which observer is measuring the proper time between the two events? Explain your answer. (2)

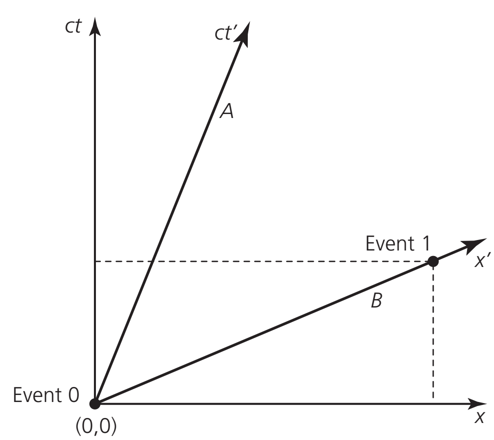
c Explain why the time interval Δt is different from the time interval $\Delta t'$. (3)

d The angle between the muon's worldline and the *x* axis is 75°. Show that the speed of the muon according to an observer at rest in reference frame S is 0.27 *c*. (2)

e The muon exists for 2.2 μs according to the observer in reference frame S. How long a time interval does the muon exist for according to the observer in reference frame S′? (3)

f The muon decays into an electron, a neutrino and an antineutrino. In this experiment the electron is virtually stationary. Draw the worldlines of the neutrino and the antineutrino if they travel in opposite directions along the *x*′ dimension at virtually the speed of light. (3)

Q5 This question is about spacetime diagrams and relativistic velocity transformations. It refers to the spacetime diagram below, which draws the reference frames A and B on the same diagram.



a What is meant by spacetime? (1)

b Are events 0 and 1 simultaneous? Explain your answer. (3)

c By taking measurements from the diagram, estimate the speed of B according to A. (3)

d A third observer, C, is also present at event 0. According to observer B, C moves after A with a velocity of 0.300 *c*. Using your answer to **b**, calculate C's speed according to observer A. (3)

e Draw the worldline for C onto the spacetime diagram. (3)

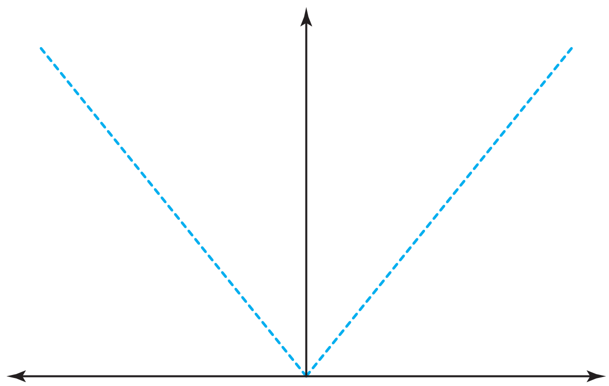
Higher level only

Q6 This question is about protons travelling around a circular particle ring accelerator. Protons are accelerated from rest through a voltage of 300 000 kV before being injected into a particle accelerator ring.

a Calculate the speed of the protons as they are injected into the ring. (3)

b Show that the momentum of the protons is approximately 0.8 GeV c^{-1}. (3)

c The protons are then accelerated in two beams travelling in opposite directions around the ring and are accelerated up to a speed of 0.97 c before colliding. On the axes below sketch the worldline of one of the protons as it orbits the accelerator ring at 0.97 c drawn from the reference frame of the laboratory. The dotted lines represent lightlines. (3)



d Calculate the total energy available when two protons travelling in opposite directions collide. (3)

e According to an observer in the reference frame of one of the protons, how fast is the other proton approaching? (3)

Q7 This question is about relativistic energy and momentum.

a Particle A is at rest with respect to an observer. Another identical particle, B, is moving with respect to the observer. Distinguish between the total energy of particle A and the total energy of particle B as measured by the observer. (2)

b Two protons are travelling towards each other along the same straight line in a vacuum. The speed of each proton, as measured in the laboratory frame of reference, is 0.960 c.

i Calculate the relative speed of one proton with respect to the other proton. (2)

ii Show that the total energy of one of the protons, according to an observer at rest in the laboratory, is 3.35 GeV. (2)

c The collision of the two protons results in the following reaction:

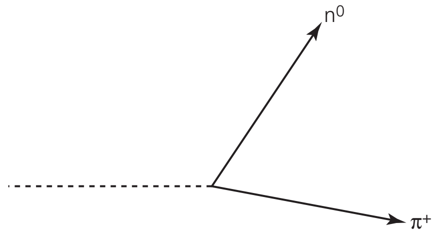
$$p^+ + p^+ = p^+ + n^0 + \pi^+$$

where π^+ is a particle called a pion and has a rest mass of 140 MeV c^{-2}. The total energy of the pion is 502 MeV. Determine, according to an observer at rest in the laboratory:

i the total energy of the proton formed plus the total energy of the neutron formed by the collision (2)

ii the momentum of the pion. (2)

d The diagram shows the paths followed by the neutron and pion in **c**.
The dotted line shows the path of the original collision of the protons in **b**. On the diagram, draw the direction of the proton formed in the collision. (1)



Q8 This question is about time dilation of GPS satellites. A GPS satellite travels in a circular orbit around the Earth with an orbital radius of 20.4×10^3 km so that it is travelling at $4.42\,\text{km s}^{-1}$. ($M_{\text{Earth}} = 5.98 \times 10^{24}$ kg, $r_{\text{Earth}} = 6.37 \times 10^6$ m.)

a The clocks in the GPS satellites experience time dilation relative to the clocks on the Earth's surface. Describe what is meant by the term *time dilation*. (2)

b Calculate:

i the fraction of the speed of light that the satellite is travelling (1)

ii the Lorentz factor at this speed. (2)

c Using Lorentz transformations, show that the time recorded by the satellite due to its relative motion will run approximately 7×10^{-6} s slow every 24.0 hours. (2)

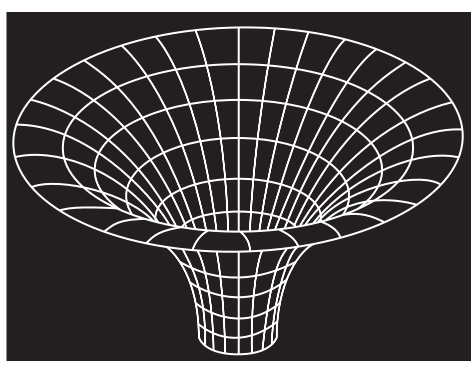
d With the aid of a diagram describe and explain what is meant by the term 'gravitational blueshift'. (3)

e According to general relativity the clocks in the GPS satellites run faster than clocks on the Earth's surface.

i Assume that $g = 9.81$ ms^{-1}. Use the gravitational redshift equation to estimate the difference in the time measured by the GPS satellites after 1 second measured on Earth. (3)

ii In 24 hours, by how much do clocks in the GPS satellites run fast relative to those on Earth, due to general relativistic effects? (1)

Q9 The diagram below illustrates the distortion of space by the gravitational field of a black hole.



a i Describe what is meant by the *centre* and the *surface* of a black hole. (3)

ii With reference to your answer in **a i**, define the 'Schwarzschild radius'. (1)

iii Calculate the Schwarzschild radius for an object having a mass of 2.0×10^{31} kg (ten solar masses). (2)

iv Science fiction frequently portrays black holes as objects that 'swallow' everything in the universe. A spacecraft is travelling towards the object in **a iii** such that, if it continues in a straight line, its distance of closest approach would be about 10^7 m. By reference to the diagram and your answer in **a iii**, suggest whether the fate of the spacecraft is consistent with the fate as portrayed in science fiction. (2)

b In 1979, Walsh, Carswell and Weymann discovered 'two' very distant quasars separated by a small angle. Spectroscopic examination of the images showed that they were identical. Outline how these observations give support to the theory of general relativity. (2)

Option B

14 Engineering physics

ESSENTIAL IDEAS

- The basic laws of mechanics have an extension when equivalent principles are applied to rotation. Actual objects have dimensions and they require an expansion of the point-particle model to consider the possibility of different points on an object having different states of motion and/or different velocities.
- The first law of thermodynamics relates the change in the internal energy of a system to the energy transferred and the work done. The entropy of the universe tends to a maximum.
- Fluids cannot be modelled as point particles. Their distinguishable response to compression from solids creates a set of characteristics that require an in-depth study.
- In the real world, damping occurs in oscillators and has implications that need to be considered.

14.1 (B1: Core) Rigid bodies and rotational dynamics

– the basic laws of mechanics have an extension when equivalent principles are applied to rotation; actual objects have dimensions and they require an expansion of the point-particle model to consider the possibility of different points on an object having different states of motion and/or different velocities

In Chapter 2 we discussed *translational* motion in which objects have their linear velocities changed by the action of resultant forces, resulting in accelerations and changes of position. Newton's three laws of motion were used to explain and predict the movement of masses in terms of the forces acting on them.

In this section we will consider how similar ideas can be applied to objects that are able to rotate. A **rotation** is the circular movement of an object around a point or an axis. If the centre of rotation is within the object, the object can be described as *spinning*. If the axis is outside the object, the object can be described as *revolving*. See Figure 14.1. Observations and measurements made on fairground rides provide interesting data for use in this section.

Figure 14.1 Many fairground rides involve spinning and revolving

Rotational dynamics is the name given to the branch of physics and engineering that deals with the motion of rotating objects. In this chapter we will discuss only **rigid bodies** – objects that keep their shape.

This topic has many applications including engines, electrical generators and motors (most involve continuous rotations including those transferred to easily observed motion in wheels, vehicle propellers and wind generators), balls, planets, galaxies and stars.

Torque

In any situation where rotation may be possible it is important to first identify the place about which the rotation can occur. Most commonly this will involve an **axis of rotation**. (The terms *pivot*, *hinge* and *fulcrum* are also widely used for various situations in which rotation is not complete or continuous.)

The **line of action** of a force is a straight line showing the direction in which the force is applied. Any force applied to an object whose line of action is *not* through the axis of rotation will tend to start or change rotational motion, if that is possible.

Clearly, bigger forces will tend to produce higher accelerations but the line of action (direction) of the force is also very important. See Figure 14.2.

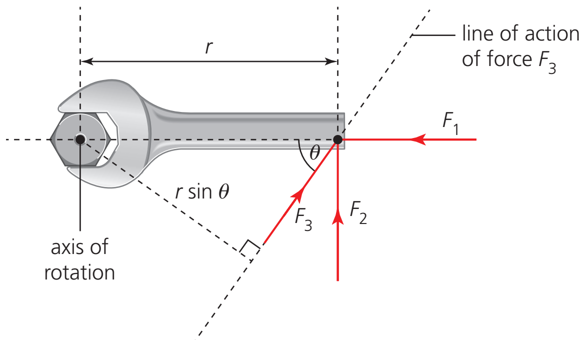


Figure 14.2 Forces of equal magnitude but different directions

Of the three forces shown in Figure 14.2, F_1 has no turning effect because its line of action is through the axis of rotation. F_2 has the biggest turning effect because its line of action is perpendicular to a line joining its point of application to the axis. F_3 has an effect between these two extremes. The turning effect also depends on the distance, r, from the axis of rotation to the line of action of the force.

In general, the 'turning effect' of a force, F, is known as its **torque**, Γ, and it depends on the magnitude of the force and the perpendicular distance from the axis of rotation to the line of action of the force (sometimes called the *lever arm*). In Figure 14.2 this is shown for force F_3 as $r \sin \theta$, where θ is the angle between the line of action of the force and a line joining the point of application of the force to the axis of rotation; r is the distance from the axis of rotation to the point of application of the force. Torque is defined as follows:

$$\Gamma = Fr \sin \theta$$

This equation is given in the *Physics data booklet*.

In some circumstances, torque is also called the **moment** of a force. Torque has the unit N m (not N m^{-1}), but note that it is *not* equivalent to the unit of energy, the joule, which is also N m.

Worked example

1 Consider Figure 14.2.

a If $r = 48$ cm, what torque is produced by a force of 35 N applied along the line of action of F_2?

b What value of F_3 would produce the same torque as in (a) if the angle $\theta = 37°$?

a $\Gamma = Fr \sin \theta$

$= 35 \times 0.48 \times \sin 90°$

$= 17$ N m

b $17 = F_3 \times 0.48 \times \sin 37°$

$$F_3 = \frac{17}{0.29} = 59 \text{ N}$$

Combining torques

Torque is a vector quantity, although generally we will only be concerned as to whether a torque tends to produce clockwise or anticlockwise motion.

When more than one torque acts on a body the resultant torque can be found by addition, but clockwise and anticlockwise torques will oppose each other and the 'direction' of torques must be taken into account when determining the resultant. (This is sometimes expressed as the **principle of moments**.)

Couples

A *couple* is the name given to a pair of equal-sized forces that have different lines of action but which are parallel to each other and act in opposite directions, tending to cause rotation.

A couple produces no resultant force on an object, so there is no translational acceleration; it will remain in the same location. Figure 14.3 shows a typical example, a couple used to turn a wheel. Other examples of using couples are the forces on a tap being turned, the forces on a bar magnet placed in a magnetic field, the forces on the pedals of a bicycle and the forces on a spinning motor.

The magnitude of the torque provided by a couple is simply twice the magnitude of the torque provided by each of the two individual forces, $\Gamma = 2Fr\sin\theta$. See Figure 14.4.

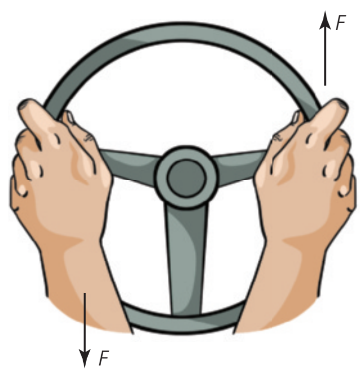


Figure 14.3 A couple used to turn a steering wheel

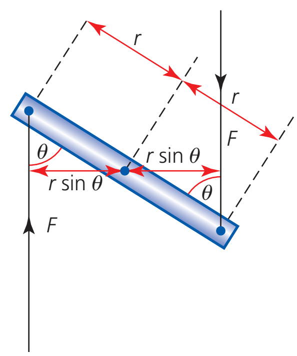


Figure 14.4 Calculating the torque provided by a couple

If a single force acts on an object (which is free to move in any direction) but not through its centre of mass, it will tend to change both its rotational and translational motion.

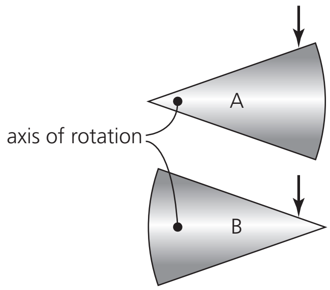
1 To loosen a nut on an engine a torque of 48 N m is required. What is the minimum force with which this can be achieved if the length of the spanner (wrench) used is 23 cm?

2 **a** Determine the torque provided by the couple shown in Figure 14.4 if the force is 42 N, $r = 8.3$ cm and the angle θ is 53°.
b Sketch a graph to show how the torque varies when the object moves from horizontal to vertical (assume that the directions of the forces do not change).

3 Consider Figure 14.4. Does the magnitude of the torque provided by the couple depend on the position of the axis of rotation? Explain your answer.

Moment of inertia

Inertia is that property of an object that resists changes of motion (accelerations). For *translational* motion, the inertia of an object depends only on its mass (the object is assumed to be free of all other forces, such as friction). In other words, larger masses require bigger forces (than smaller masses) to produce equal accelerations. We say that larger masses have higher inertia, and that mass can be considered as a measure of inertia. Newton's first and second laws can be expressed in terms of inertia, but the concept of inertia was not included in Chapter 2 when Newton's laws were introduced.



■ **Figure 14.5** Different axes can produce different moments of inertia for the same object

Resistance to *rotational* motion also depends on the mass of the object, and also on how the mass is distributed about the axis of rotation. Consider Figure 14.5. Object A will require more force to produce a certain acceleration than object B, which has the same mass and shape but a different axis of rotation.

Resistance to a change of rotational motion of an object is quantified by its **moment of inertia**, I, which depends on the distribution of mass around the chosen axis of rotation.

The simplest object to consider is a *point mass*. The moment of inertia of a point mass, m, rotating at a distance r from its axis of rotation is given by:

$$I = mr^2$$

The unit of moment of inertia is kg m^2. Most spherical objects can be considered to behave like masses concentrated at their centre points.

Nature of Science

Point particles

A *point particle* (also sometimes called an 'ideal' particle) is a concept used widely in physics, such as in work on gravitational and electric fields. A point particle does not occupy any space and whenever we refer to a real mass as behaving like a point mass, we are suggesting that we do not need to be concerned about its actual shape or size. For example, a mass that behaves as a point mass has properties that depend only on the magnitude of that mass.

The concept of a point particle simplifies the physics involved in any theory; such an assumption may be reasonable for a particle that is a relatively long way away, and it may be valid for spherical, homogeneous particles under most circumstances. Even if the assumption is not reasonable, the basic theory can be adapted to suit individual circumstances.

Worked example

2 What is the moment of inertia of a 10 g simple pendulum of length 75 cm?

$I = mr^2$

$= 0.01 \times 0.75^2$

$= 5.6 \times 10^{-3}$ kg m^2 (assuming the pendulum can be considered to act as a point mass)

The moment of inertia of any real, extended mass can be determined by the addition of the individual moments of inertia of its particles. This is represented by:

$$I = \Sigma mr^2$$

This equation is given in the *Physics data booklet* (the symbol Σ means 'sum of').

This equation can be used to determine the moment of inertia of various different shapes. Some examples are given below, but there is no need to remember them or know how they were derived because equations will be provided in the examination paper if needed:

- solid sphere of radius r about an axis through its centre: $\frac{2}{5}mr^2$
- solid cylinder or disc of radius r with a central axis (parallel to sides): $\frac{1}{2}mr^2$
- thin-walled hollow cylinder of radius r with a central axis (parallel to sides): mr^2
- thin rod of length L, axis perpendicular to rod through centre: $\frac{1}{12}mL^2$
- thin rod of length L, axis perpendicular to rod through end: $\frac{1}{3}mL^2$

Utilizations

Flywheels

Flywheels are designed to have large moments of inertia. They are added to the axes of rotating machinery to resist changes of motion and/or to store rotational kinetic energy. They are used in modern machinery, but Figure 14.6 shows an old-fashioned example, a potter's wheel. The large wheel at the bottom is kicked for a while until the system is spinning quickly. After that, because it has a large moment of inertia, there will be no need to keep kicking the wheel continuously.

■ **Figure 14.6** A flywheel on a potter's wheel

Flywheels can be useful for maintaining rotations in machines that do not have continuous power supplies. To do this they will usually need to be able to store relatively large amounts of kinetic energy.

1 Estimate the kinetic energy stored in a flywheel of mass 100 kg and radius 30 cm if it is spinning at a rate of 50 revolutions every second. (Use the equation $E_K = \frac{1}{2}I\omega^2$, which is explained later in this section.)

4 For the five different moments of inertia listed above, draw diagrams to represent their shapes and axes of rotation.

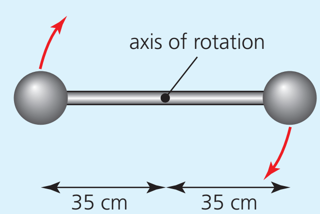
5 Estimate the moment of inertia of the Earth in its orbit around the Sun (mass of Earth ≈ 6×10^{24} kg, distance to Sun ≈ 150 Mkm).

6 **a** What is the moment of inertia of a spinning golf ball of mass 46 g and radius 21 mm?
b What assumption did you make in answering (a)?

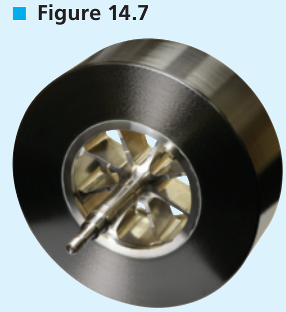
7 **a** Suggest why the equation $I = mr^2$ can be used to approximate the moment of inertia of a bicycle wheel.
b Calculate a value for the moment of inertia of a typical bicycle wheel by first estimating its mass and radius.

8 Figure 14.7 shows an arrangement in which two spherical masses, each of mass 2.0 kg are rotating about an axis that is 35 cm from both of their centres.
a If the rod has a mass of 400 g and a length of 56 cm, determine the overall moment of inertia of this arrangement.
b What percentage does the rod contribute to the overall moment of inertia of the system?

9 The flywheel shown in Figure 14.8 may be considered to be a steel cylinder (density 7800 kg m^{-3}) of outer radius 40 cm and depth 12 cm, but with the central section (of outer radius 18 cm) missing.
a Use this simplification to estimate its moment of inertia. Ignore the supporting structure.
b Research any one practical application of a flywheel and explain in detail *why* the flywheel is used.



■ **Figure 14.7**



■ **Figure 14.8** Flywheel

Rotational and translational equilibrium

If an object remains at rest or continues to move in exactly the same way, it is described as being in *equilibrium*.

Translational equilibrium occurs when there is no resultant force acting on an object, so that it remains stationary or continues to move with a constant velocity (that is, in a straight line at a constant speed).

Rotational equilibrium occurs when there is no resultant *torque* acting on an object, so that it remains stationary or continues to rotate with a constant angular velocity.

Angular acceleration

Remember from Chapter 6 that **angular velocity** of an object rotating uniformly can be calculated from:

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

This equation is given in the *Physics data booklet*. The unit of angular velocity is $\mathrm{rad\,s^{-1}}$.

If an object experiences a resultant torque, it will have an angular acceleration (a change to its angular velocity).

Angular acceleration, α, is defined as the rate of change of angular velocity with time:

$$\alpha = \frac{\Delta\omega}{\Delta t}$$

So that, when an angular velocity changes from ω_i to ω_f in a time, t, the angular acceleration, α, is given by:

$$\alpha = \frac{\omega_f - \omega_i}{t}$$

or:

$$\omega_f = \omega_i + \alpha t$$

This equation is given in the *Physics data booklet*. It is comparable to $v = u + at$ for translational motion.

Angular acceleration has the unit $\mathrm{rad\,s^{-2}}$.

Relating angular acceleration to linear acceleration

We know from Chapter 6 that:

$$v = \omega r$$

so that:

$$\alpha = \frac{\Delta\omega}{\Delta t} = \frac{\Delta v}{r\Delta t}$$

or:

$$\alpha = \frac{a}{r}$$

This equation is not given in the *Physics data booklet*.

Equations of rotational motion for uniform angular acceleration

The following equations were used in Chapter 2 to describe translational motion at constant linear acceleration:

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

$$v^2 = u^2 + 2as$$

We have already noted similarities between the concepts and equations used in linear and rotational mechanics, and this can be continued by using the **analogy** to write down two further equations that describe rotational motion with *constant angular acceleration*:

$$\theta = \omega_i t + \frac{1}{2}\alpha t^2$$

$$\omega_f^2 = \omega_i^2 + 2\alpha\theta$$

These equations are given in the *Physics data booklet*.

10 The circumference of the London Eye (Figure 14.9) can rotate continuously at a speed of 26 cm s^{-1}. Its radius is 60 m.
- **a** Calculate its angular velocity.
- **b** Calculate how many minutes it takes to complete one revolution.

11 The outer rim of a bicycle wheel of radius 32 cm has a linear acceleration of 0.46 m s^{-2}.
- **a** What is the angular acceleration of the wheel?
- **b** If it starts from rest, what time is needed for the wheel to accelerate to a rate of three revolutions every second?

12 A wheel accelerates for 5.0 s from rest at 5.2 rad s^{-2}.
- **a** What is its angular velocity at the end of 5 s?
- **b** What is its total angular displacement in this time?
- **c** How many rotations does it complete in 5 s?
- **d** After 5.0 s the accelerating torque is removed and the wheel decelerates at a constant rate to become stationary again after 18.2 s. How many rotations are completed during this time?

Figure 14.9 The London Eye

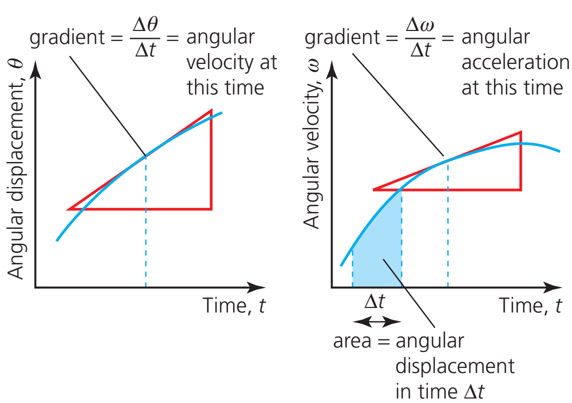
13 A rotating fan blade has an angular velocity of 7.4 rad s^{-1}. It is then made to accelerate for 1.8 s, during which time it passes through an angular displacement of 26.1 rad. Calculate the angular acceleration of the fan blade.

14 A machine spinning at 3000 revolutions per minute (rpm) is accelerated to 6000 rpm while the machine made 12 revolutions.
- **a** Convert 3000 rpm to rad s^{-1}.
- **b** What was the angular acceleration?

Graphs of rotational motion

We can also use analogies with linear mechanics when interpreting graphs of rotational motion. See Figure 14.10.

Figure 14.10 Interpreting graphs of rotational motion



Newton's second law applied to angular motion

For translational motion, Newton's second law shows that when a resultant force, *F*, acts on an object of mass, *m*, it produces a linear acceleration, *a*, given by:

$$F = ma$$

Similar ideas can be applied to *rotational* motion – when a resultant torque, Γ, acts on an object with moment of inertia, I, it produces an angular acceleration, α:

$$\Gamma = I\alpha$$

This equation is given in the *Physics data booklet*.

Worked example

3 **a** What acceleration is produced when a system that has a moment of inertia of 0.54 kg m^2 is acted on by a resultant torque of 0.76 N m?
b If the system was already rotating at 1.7 rad s^{-1}, what is its maximum angular velocity if the torque is applied for exactly 3 s?
c What assumption did you make when answering these questions?

a $\Gamma = I\alpha$

$\alpha = \frac{\Gamma}{I} = \frac{0.76}{0.54} = 1.4\,\text{rad s}^{-2}$

b $\omega_f = \omega_i + \alpha t = 1.7 + (1.4 \times 3.0) = 5.9\,\text{rad s}^{-1}$

c There are no frictional forces acting on the system.

This equation can be used in an experimental determination of moment of inertia. For example, see Figure 14.11 (and Figure 14.12). The moment of inertia of the system can be calculated if the angular acceleration produced by a known force can be determined.

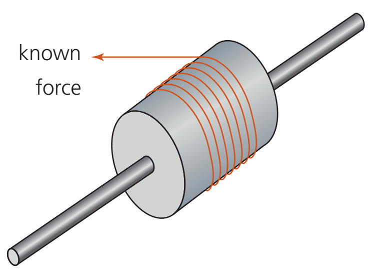


Figure 14.11 Rotating a cylinder with a constant force

Worked example

4 Figure 14.12 shows a falling mass, m, attached to a string which is wrapped around a cylinder of radius r and moment of inertia $\frac{1}{2}Mr^2$. Derive an equation for the downward acceleration of the mass.

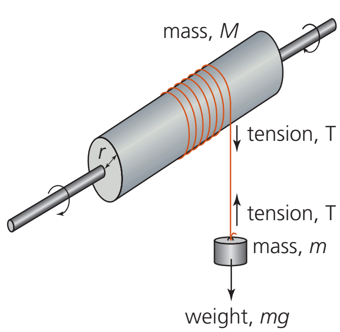


Figure 14.12

torque acting on cylinder, $\Gamma = Tr = I\alpha = \frac{1}{2}Mr^2 \times \frac{a}{r}$

$T = \frac{1}{2} \times \frac{Mr^2}{r} \times \frac{a}{r} = \frac{1}{2}Ma$

resultant force acting on falling mass, $F = mg - T$

linear acceleration of falling mass, $a = \frac{F}{m} = \frac{mg - T}{m}$

$= \frac{mg - \frac{1}{2}Ma}{m}$

Rearranging gives:

$a = \frac{mg}{m + \frac{1}{2}M}$

15 Calculate the moment of inertia of a system that is accelerated from 500 rpm (revolutions per minute) to 1500 rpm in 1.3 s when a resultant torque of 98 N m is applied.

16 Consider Figure 14.11. When a constant force of 25 N was applied for 2.0 s the rotating cylinder accelerated from rest to pass through a total angle of 22.4 radians.
a What was the angular acceleration?
b Determine the moment of inertia of the system if the radius of the cylinder was 4.4 cm.

17 A couple is provided by two parallel forces each of 26 N separated by a distance of 8.7 cm. If this couple provides the resultant torque to a rotating system that has a moment of inertia of 17.3 kg m^2, determine the angular acceleration produced.

18 A torque of 14.0 N m is applied to a stationary wheel, but resistive forces provide an opposing torque of 6.1 N m. Calculate the angular velocity after 2.0 s if the wheel has a moment of inertia of 1.2 kg m^2.

19 Car manufacturers usually quote the output power of their vehicles and the maximum torque provided. Discuss how these two quantities are related to each other.

20 Sketch an angular displacement graph for the following rotational motion: an object rotates at 3 rad s^{-1} for 4 s, it then very rapidly decelerates and then remains stationary for a further 6 s. The rotation is then reversed so that it accelerates back to its original position after a total time of 15 s.

21 Figure 14.13 shows how the angular velocity of an object changed during 6 s.
a What was the angular acceleration during the first 2 s?
b Through what total angle did the object rotate in 6 s?

22 Consider Figure 14.12. What mass, *m*, will have a linear acceleration of 2.5 m s^{-2} when acting on an 8.3 kg cylinder?

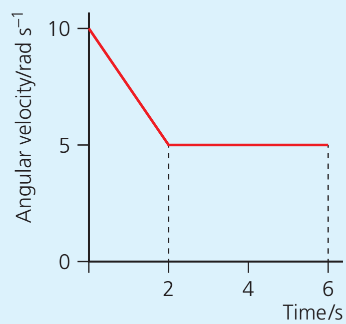


Figure 14.13

Conservation of angular momentum

The **angular momentum**, *L*, of a rotating object is the rotational equivalent of linear momentum. It depends on the moment of inertia, *I*, of the object and its angular velocity, ω. It is defined as follows:

$$L = I\omega$$

This equation is given in the *Physics data booklet*. The unit of angular momentum is kg m^2 s^{-1}.

Worked example

5 A sphere of mass 2.1 kg and radius 38 cm is spinning at a rate of 44 rpm. What is its angular momentum?

$$L = I\omega = \frac{2}{5}mr^2 \times \frac{2\pi}{T} = \frac{2}{5} \times 2.1 \times 0.38^2 \times 2 \times \frac{\pi}{60/44}$$
$$= 0.56\ \text{kg m}^2\,\text{s}^{-1}$$

In Chapter 2 the law of conservation of linear momentum, which has no exceptions, was shown to be very useful when predicting the outcome of interactions between masses exerting forces on each other. In a similar way, the law of conservation of angular momentum (as follows) has no exceptions and can be used to predict changes to rotating systems.

The total angular momentum of a system is constant provided that no external torques are acting on it.

For example, figure skaters (Figure 14.14), gymnasts and ballet dancers can spin faster by moving their arms and legs closer to their body's axis of rotation. This decreases their moment of inertia so, because their angular momentum is constant (if we assume no significant external forces are acting), their angular velocity must increase.

Figure 14.14 Spinning figure skater

Worked example

6 A solid metal disc of mass 960 g and radius 8.8 cm is rotating at 4.7 rad s^{-1}.

a Calculate the moment of inertia of the disc.

b Calculate the new angular velocity after a mass of 500 g is dropped quickly and carefully on to the disc at a distance of 6.0 cm from the centre.

a $I = \frac{1}{2}mr^2 = \frac{1}{2} \times 0.96 \times (8.8 \times 10^{-2})^2$

$= 3.7 \times 10^{-3}\,\text{kg m}^2$

b Moment of inertia of added mass $= mr^2 = 0.5 \times (6.0 \times 10^{-2})^2$

$= 1.8 \times 10^{-3}\,\text{kg m}^2$

$L = I\omega = \text{constant}$

$(3.7 \times 10^{-3}) \times 4.7 = ((3.7 \times 10^{-3}) + (1.8 \times 10^{-3})) \times \omega$

$\omega = 3.2\,\text{rad s}^{-1}$

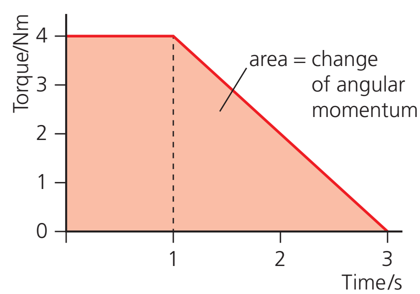


Figure 14.15 Example of a torque–time graph

Torque–time graphs

You know from Chapter 2 that the area under a force–time graph is equal to the change of momentum of the system (impulse). Similarly, in rotational dynamics the area under a torque–time graph is equal to the change in angular momentum. This is true for any shape of graph. Consider Figure 14.15.

Area under graph = change of angular momentum $= (4.0 \times 1.0) + \frac{1}{2}(4.0 \times 2.0) = 8.0\,\text{N m s}$ (or kg $m^2\,s^{-1}$).

23 An unpowered merry-go-round of radius 4.0 m and moment of inertia 1200 kg m^2 is rotating with a constant angular velocity of 0.56 rad s^{-1}. A child of mass 36 kg is standing close to the merry-go-round and decides to jump onto its edge.

a Calculate the new angular velocity of the merry-go-round.

b Discuss whether the merry-go-round would return to its original speed if the child jumped off again.

24 Neutron stars are the very dense collapsed remnants of much larger stars. Suggest why they have extremely high rotational velocities.

25 Figure 14.16 shows the variation of torque applied to a stationary system that has a moment of inertia of 0.68 kg m^2.

a Estimate the change of angular momentum of the system.

b What is its final angular velocity?

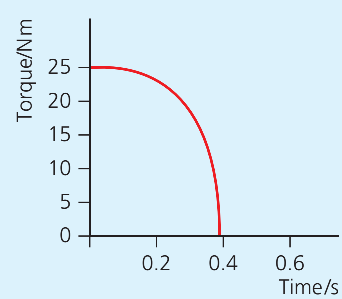


Figure 14.16

Rotational kinetic energy

Translational kinetic energy can be calculated from the equation $\frac{1}{2}mv^2$. By analogy, rotational kinetic energy can be calculated from:

$$E_{K_{rot}} = \frac{1}{2}I\omega^2$$

This equation is given in the *Physics data booklet*.

It is common for objects to have both rotational and translational kinetic energy at the same time – the wheels on a moving vehicle are a good example. Gas molecules also have both kinds of kinetic energy.

Worked example

7 Calculate the rotational kinetic energy of a tossed coin if it has a mass of 8.7 g, radius 7.1 mm and completes one rotation in 0.52 s.

$$\omega = 2\pi f = 2\pi \times \frac{1}{0.52} = 12\,\text{rad s}^{-1}$$

$$I = \frac{1}{2}mr^2 = \frac{1}{2} \times (8.7 \times 10^{-3}) \times (7.1 \times 10^{-3})^2 = 2.2 \times 10^{-7}\,\text{kg m}^2$$

$$E_{K_{rot}} = \frac{1}{2}I\omega^2 = \frac{1}{2} \times (2.2 \times 10^{-7}) \times 12^2$$

$$= 1.6 \times 10^{-5}\,\text{J}$$

The moment of inertia of an object (such as a wheel) mounted on an axle can be determined experimentally as shown in Figure 14.17. As discussed earlier, this is a simple experiment if a way of providing a constant force is available. If not, a conservation of energy approach can be helpful, as follows.

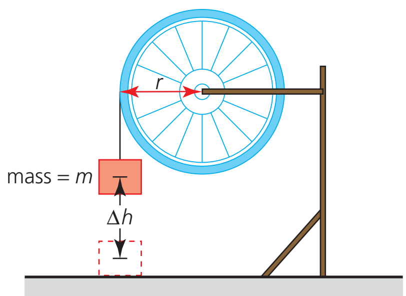


Figure 14.17 Determining the moment of inertia of a wheel

In Figure 14.17 a string is wrapped around the outside of a wheel and provides a torque as the attached mass falls and starts the wheel rotating. The angular velocity of the wheel must be measured when the mass has fallen a known distance. Videoing this kind of motion experiment can be very helpful for obtaining data.

After the mass has fallen a distance, Δh, it has transferred gravitational potential energy, $mg\Delta h$, to kinetic energy of itself and of the rotating wheel.

$$mg\Delta h = \frac{1}{2}mv^2 + \frac{1}{2}I\omega^2$$

Rolling

One very common kind of rotation is that of a rolling object, such as a wheel. Rolling is a combination of translational and rotational motion. Without friction, rolling is not possible on a horizontal surface – an object can only slip and slide. Here we shall only consider rolling *without* slipping.

Figure 14.18 shows vector arrows representing the velocities of certain locations on a wheel.

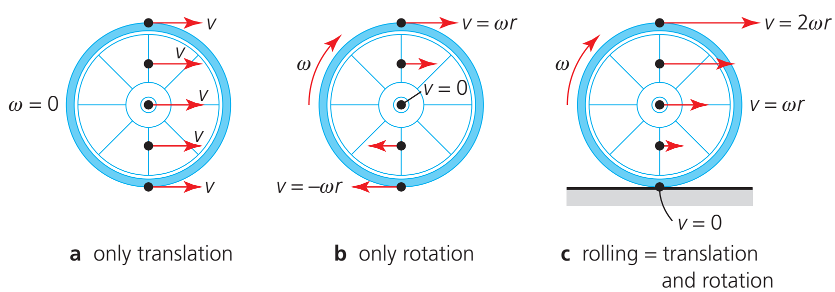


Figure 14.18 Different motions of a wheel (of radius *r*)

- In (a) the wheel shown is *not* rotating but it is moving to the right with translational motion. In this unusual situation, all parts of the wheel have the same *translational linear* velocity at the same time.
- In (b) the wheel is rotating but *not* changing its position. Perhaps there is no friction or the wheel is not touching the ground. There is no translational motion. All parts of the wheel have the same *angular* velocity, ω, but any instantaneous *linear* velocity, v, depends on the location of the point on the wheel ($v = \omega r$).
- In (c) the wheel is rotating *and* moving to the right at the same time. The translational *linear* velocity of the wheel is equal to the speed of a point on the circumference ($v = \omega r$).

Because we are assuming that there is no slipping of the wheel on the surface when it touches the ground, the lowest point of the wheel must be instantaneously stationary:

- ☐ velocity of bottom of wheel = linear velocity to right, $+\omega r$, + velocity of a point at the bottom of the wheel, $-\omega r$, = 0
- ☐ velocity at the top of the wheel = linear velocity to right, $+\omega r$, + velocity of a point at the top of the wheel, $+\omega r$, = $2\omega r$

Rolling down a slope

An object, such as a ball or a wheel, that can roll down a hill will transfer its gravitational potential energy to both translational kinetic energy and rotational kinetic energy. See Figure 14.19.

$$mg\Delta h = \frac{1}{2}mv^2 + \frac{1}{2}I\omega^2$$

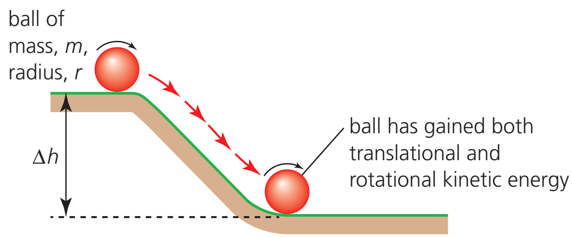


Figure 14.19 Rolling down a slope

This means that, at the bottom of a friction-free slope, a sliding object would reach a higher speed than a rolling object. Also, rotating objects that have bigger moments of inertia will travel slower at the bottom of the same slope. If the angle of the slope is too steep, rolling will not be possible. These variations can provide the basis for laboratory investigations.

Consider the example of a solid sphere, for which $I = \left(\frac{2}{5}\right)mr^2$, remembering that $v = \omega r$, the equation above becomes:

$$\cancel{m}g\Delta h = \frac{1}{2}\cancel{m}\omega^2 r^2 + \frac{1}{2}\left(\frac{2}{5}\right)\cancel{m}r^2\omega^2 = \frac{7}{10}\omega^2 r^2$$

Note that, with the assumptions made, the angular velocity at the bottom of the slope does not depend on the slope angle or the mass of the ball.

Comparison of concepts and equations between linear and rotational mechanics

Table 14.1 summarizes all the rotational mechanics concepts and equations covered in this section, and compares them to the analogous concepts in linear mechanics covered in Chapter 2. The bold items are provided in the *Physics data booklet*.

Table 14.1 Summary of the concepts and equations for linear and rotational mechanics

Linear mechanics	Rotational mechanics
displacement, s initial velocity, $u = \frac{\Delta s}{\Delta t}$ final velocity, v acceleration, $a = \frac{\Delta v}{\Delta t}$	angular displacement, θ initial angular velocity, $\omega_i = \frac{\Delta\theta}{\Delta t}$ final angular velocity, ω_f angular acceleration, $\alpha = \frac{\Delta\omega}{\Delta t}$
$v = u + at$ $s = \left(\frac{u+v}{2}\right)t$ $s = ut + \frac{1}{2}at^2$ $v^2 = u^2 + 2as$	angular velocity, $\omega = \frac{2\pi}{T} = 2\pi f$ $\boldsymbol{\omega_f = \omega_i + \alpha t}$ $\theta = \left(\frac{\omega_i + \omega_f}{2}\right)t$ $\boldsymbol{\theta = \omega_i t + \frac{1}{2}\alpha t^2}$ $\boldsymbol{\omega_f^2 = \omega_i^2 + 2\alpha\theta}$
force, F mass, m $F = ma$	torque, $\boldsymbol{\Gamma = Fr\sin\theta}$ moment of inertia, $\boldsymbol{I = \sum mr^2}$ $\boldsymbol{\Gamma = I\alpha}$
linear momentum, $p = mv$	angular momentum, $\boldsymbol{L = I\omega}$
momentum is always conserved in all interactions provided that no external forces are acting.	
linear kinetic energy, $E_K = \frac{1}{2}mv^2$	rotational kinetic energy, $\boldsymbol{E_{K_{rot}} = \frac{1}{2}I\omega^2}$

26 The moment of inertia of a domestic cooling fan is 8.2×10^{-4} kg m^2. If the fan rotates at a frequency of 20 Hz, calculate its rotational kinetic energy.

27 A car travelling at a constant velocity of 23 m s^{-1} has a wheel with a radius 32 cm and mass of 8.9 kg.
- **a** Determine the angular velocity of the wheel.
- **b** If the wheel has a moment of inertia of 0.72 kg m^2, calculate its rotational kinetic energy.
- **c** Compare your answer to (b) to the translational kinetic energy of the wheel at the same speed.

28 Consider the experiment shown in Figure 14.17. The angular velocity of the wheel after a 100 g mass had fallen 50 cm was determined to be 2.7 rad s^{-1}.
- **a** At any instant the mass was falling at the same linear speed, v, as the edge of the wheel was moving. Determine a value for v when the mass has fallen 50 cm if the radius of the wheel was 35 cm.
- **b** Calculate a value for the moment of inertia of the wheel.

29 Use the analogies between linear and rotational mechanics to write down equations for rotational work and power.

30 A solid ball and a hollow ball of the same mass and radius roll down a hill. At the bottom, which ball will be (a) rotating faster, (b) have the greater linear speed? Explain your answers.

31 Calculate the greatest angular velocity of a solid ball of radius 5.2 cm rolling down a slope of vertical height 6.0 cm. What assumption did you make?

14.2 (B2: Core) Thermodynamics *– the first law of thermodynamics relates the change in the internal energy of a system to the energy transferred and the work done; the entropy of the universe tends to a maximum*

This section is about studying the process of using thermal energy to do useful mechanical work in *heat engines* (introduced in Chapter 3). This branch of physics is known as *thermodynamics*. Although thermodynamics grew out of a need to understand heat engines, it has much wider applications. The study of thermodynamics leads to a better understanding of key scientific concepts such as internal energy, heat, temperature, work and pressure, and how they are all connected to each other and to the microscopic behaviour of particles.

Heat engines of many kinds, like those shown in Figure 3.26, play a very important part in all our lives and we will concentrate our attention on understanding the basic principles by considering those processes that involve the volume increase (*expansion*) of fixed masses of *ideal* gases.

In the rest of this chapter, and throughout physics, there are many references to (thermodynamic) 'systems and surroundings'. Before going any further we should make sure that these simple and widely used terms are clearly understood.

- A **system** is simply the thing we are studying or talking about. In this chapter it will be a gas.
- The **surroundings** are everything else – the gas container and the rest of the universe.

Sometimes we call the surroundings the *environment*. If we want to suggest that a part of the surroundings was deliberately designed for thermal energy to flow into it or out of it, we may use the term (thermal) **reservoir**.

A thermodynamic system can be as complex as a rocket engine, planet Earth or a human body, but in this chapter we will restrict ourselves to the behaviour of ideal gases in heat engines.

Additional Perspectives

Understanding terms: engine, motor or machine?

Most people would find it difficult to distinguish between a machine, a motor and an engine because, in everyday life, the three terms can sometimes be used for the same thing (e.g. in a car). Also, the definitions of words often evolve and scientific words may be absorbed more generally into everyday language – phrases such as 'search engine' and 'a motor for change' only further blur the meaning of the words. In practice, the words being discussed here have overlapping meanings.

A 'machine' is a general term meaning anything that can do something useful by transferring energy (whether it is a simple, single device or a complicated mechanism involving many parts). However, in everyday life the word is most commonly used when referring to large, complex machines that involve movement; that is, doing work. But devices such as knives and scissors are also machines in the precise sense of the word, and most physics courses teach important and basic physics ideas using simple machines such as ramps, pulleys, gears and levers. The word 'device' normally means a fairly simple machine that has been made for a specific purpose, such as a bottle opener or a hairdryer. Parts of larger machines are often called 'components'.

Figure 14.20 Internal combustion engine

'Engine' is a more specific term than 'machine'. It means a type of machine designed to make something move, but the use of the word is almost always limited to those machines that use the thermal energy (heat) obtained by burning a fuel to do work, for example an internal combustion engine in a car (Figure 14.20). To make that use clear, they are often called *heat* engines.

Figure 14.21 Rocket being launched

A 'jet engine' is a particular kind of heat engine in which the forward force (thrust) is produced by ejecting a gas or liquid in the opposite direction with high speed. The law of conservation of momentum or Newton's third law of motion are usually used to explain jet propulsion (see Chapter 2).

Most heat engines use the combustion of a fuel when it is combined in a chemical reaction with oxygen from the air outside; if there is no air available then an oxidizing agent (oxidant) must be provided from within the device, and it is then called a 'rocket' (Figure 14.21).

A 'motor' is another name for a machine that is designed to produce motion, but its most common specific use is with electric motors. However, the word has become closely linked with transport, such as motor cars and motor buses.

1 The following are all heat engines: jet, rocket, internal combustion (car), diesel, steam. Write some short notes to explain the principal differences between them.

Internal energy of an ideal gas

By definition, there are no intermolecular forces in an ideal gas (except during collisions), so the internal energy of an ideal gas is only the random kinetic energy of the molecules. In Chapter 3 you saw that the total random kinetic energy of the molecules of one mole of an ideal gas at a temperature T (K) was $\frac{3}{2}RT$, where R is the universal gas constant. For a sample of gas containing n moles, the internal energy (given the symbol U) can be determined from the equation:

$$U = \frac{3}{2}nRT$$

This equation is given in the *Physics data booklet*.

Worked example

8 **a** What is the total internal energy in one mole of an ideal gas at 0°C?
b How much energy is needed to raise the temperature of 4.8 moles of an ideal gas from 25°C to 50°C?

a $U = \frac{3}{2}nRT = \frac{3}{2} \times 1 \times 8.31 \times 273$

$= 3.4 \times 10^3$ J

b $\Delta U = \frac{3}{2}nR(323 - 298) = \frac{3}{2} \times 4.8 \times 8.31 \times 25$

$= 1.5 \times 10^3$ J

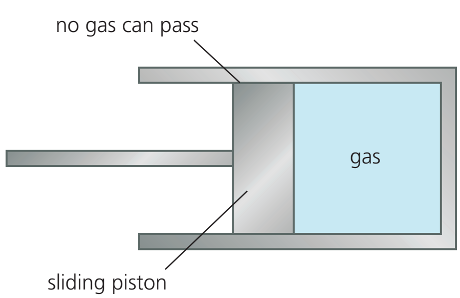


Figure 14.22 Gas in a cylinder with a movable piston

Work done when a gas changes state

For simplicity, the thermodynamic system that we are considering is often shown as a gas in a regularly shaped cylindrical container, constrained by a gas-tight **piston** that can move without friction (Figure 14.22).

First consider the idealized, simplified example of a gas expanding slightly, so that there is no change in pressure or temperature.

If the gas trapped in the cylinder in Figure 14.23 exerts a resultant force on the piston (because the pressure in the cylinder is higher than the pressure in the surroundings), the piston will move outwards as the gas expands. We say that work has been done *by* the gas in pushing back the surrounding gas (remember that we are assuming there is no friction). Because the displacement, Δs, is small, there will be no significant change in the gas pressure, p. (Both the symbols P and p are widely used to represent pressure.) Assuming the temperature is constant, we can write:

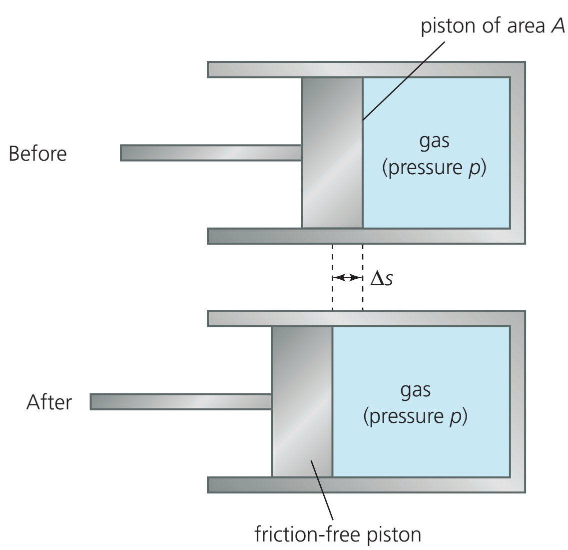


Figure 14.23 Gas expanding in a cylinder

work done by gas = force × distance moved in direction of force

or:

work done by gas = $(pA)\Delta s$ (because force = pressure × area)

Because $\Delta V = A\Delta s$:

$$W = p\Delta V$$

This equation is given in the *Physics data booklet*. If the gas is compressed, then work is done *on* the gas (system) and W is given a negative sign.

Such changes can be represented on a pV diagram (graph) as in the example shown in Figure 14.24, which shows the expansion of a gas from 0.020 m^3 to 0.050 m^3 at a constant pressure of 1.0×10^5 Pa.

The work done by the gas in expansion in this example, $W = p\Delta V$, is therefore:

$$W = (1.0 \times 10^5) \times (0.050 - 0.020) = 3.0 \times 10^3\ \text{J}$$

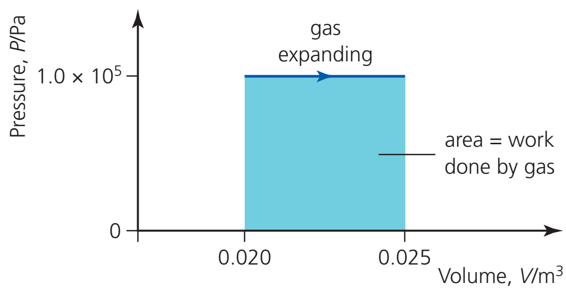


Figure 14.24 Work done during expansion of an ideal gas

Note that this calculation to determine the work done, $p\Delta V$, is numerically equal to calculating the area under the pV diagram. This is true for *all* thermodynamic processes, regardless of the shape of the graph, and this is one reason why pV diagrams are so widely used in thermodynamics to represent various processes. Figure 14.25 shows two further examples.

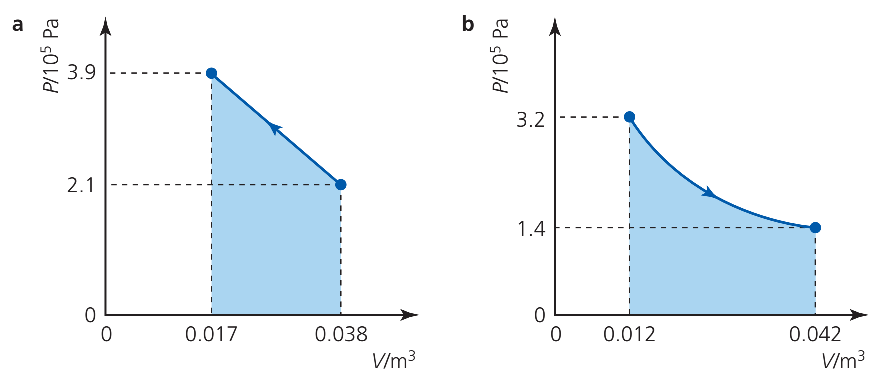


Figure 14.25 Determining areas under pressure–volume graphs

Worked example

9 Determine values for the work done in the two changes of state represented in Figure 14.25.

a $W = p\Delta V$ = area under graph $= \frac{1}{2}(3.9 - 2.1) \times 10^5 \times (0.038 - 0.017) + (0.038 - 0.017) \times 2.1$
$= 6.3 \times 10^3$ J

The work is done *on* the gas as it is compressed into a smaller volume.

b In this example, work is done *by* the gas as the volume increases. Because the graph is curved, the area underneath it must be estimated. This can be done by drawing a rectangle that is judged by eye to have the same area as that shaded in the figure:

W = area under graph $\approx (2.2 \times 10^5) \times (0.042 - 0.012)$
$= 6.6 \times 10^3$ J

First law of thermodynamics

Expansion, or any other change to the physical state of a gas, involves an energy transfer. Thermodynamic processes may involve doing mechanical work, W, changing the internal energy of the gas, U, or transferring thermal energy, Q. Remember that thermal energy (heat) is the *non-mechanical* transfer of energy from hotter to colder.

We can use the principle of conservation of energy to describe how these quantities are connected – if an amount of thermal energy, $+Q$, is transferred *into* the gas then, depending on the particular circumstances, the gas may *gain* internal energy, $+\Delta U$, and/or the gas will expand and do work *on* the surroundings, $+W$. That is:

$$Q = \Delta U + W$$

This equation is given in the *Physics data booklet*.

This important equation is known as the **first law of thermodynamics** – which is a statement of the principle of conservation of energy as applied to thermodynamic systems.

The directions of energy transfers, as shown by positive or negative signs, can cause confusion. Remember that:

- $-Q$ refers to a flow of thermal energy *out* of the gas
- $-\Delta U$ refers to a *decrease* in internal energy of the gas
- $-W$ refers to work done *on* the gas by the surroundings during compression.

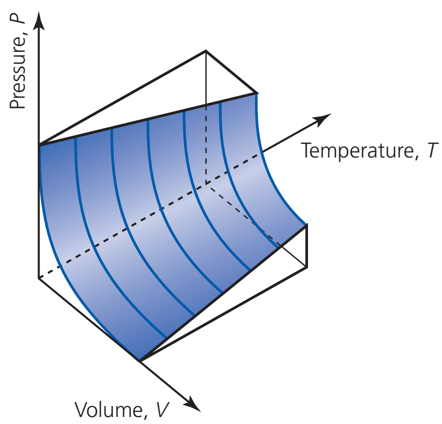


Figure 14.26 Three-dimensional curved surface representing the possible states of a fixed amount of an ideal gas

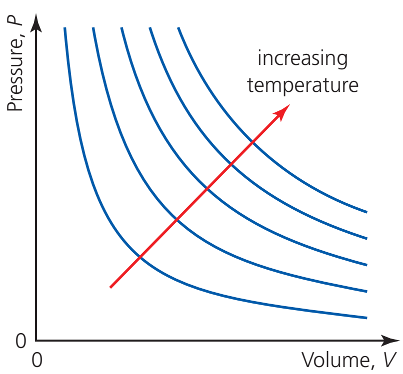


Figure 14.27 Isotherms on a *pV* diagram

pV diagrams

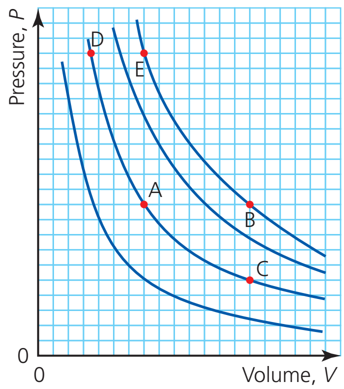
Three quantities – pressure, volume and temperature – are needed to completely specify the state of a given amount of a gas. We would need a three-dimensional graph to represent all three possibilities, as shown in Figure 14.26. Instead, we usually use a *pV diagram*, with variations in the third variable, temperature, represented by using different curved lines on the diagram, as shown in Figure 14.27.

A line on a *pV* diagram that connects possible states of the gas at the same temperature is called an **isothermal line** or **isotherm**. Several isotherms can be shown on the same *pV* plot, as in Figure 14.27. Each of these lines represents the inverse proportionality of a Boyle's law relationship, $p \propto 1/V$, as previously discussed in Chapter 3. Note that if a sample of gas is heated in the same volume to a higher temperature, then the pressure will be higher, so on *pV* diagrams 'higher' isotherms correspond to higher temperatures.

During a change of state the system will move to a new position on the *pV* diagram, but the value of $\frac{pV}{T}$ remains unaltered for a fixed amount of an ideal gas. From Chapter 3 remember:

$$pV = nRT$$

■ **Figure 14.28** Changes of state on a *pV* diagram



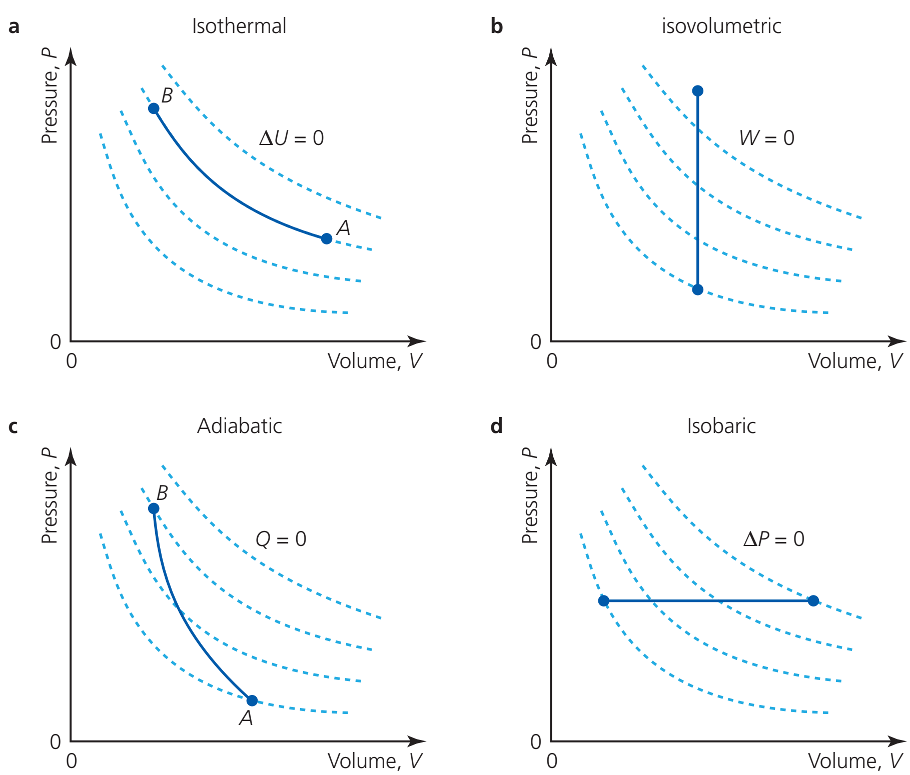
Interpreting *pV* diagrams is an important skill and Figure 14.28 shows some examples. Consider an ideal gas in a state shown by point A. Doubling the volume at constant pressure is represented by moving to point B, which is at a higher temperature. Point C also represents a doubling of volume, but at constant temperature and lower pressure. Moving to point D could be achieved by doubling the pressure at the same temperature in half the volume. Finally, point E represents doubling the pressure at a higher temperature in the same volume.

Isovolumetric, isobaric, isothermal and adiabatic processes

Among all the various changes of state that an ideal gas might experience, it is convenient to consider the first law of thermodynamics ($Q = \Delta U + W$) under four extremes (see Figure 14.29):

- $\Delta U = 0$
- $W = 0$
- $Q = 0$
- $\Delta p = 0$

■ **Figure 14.29** Four thermodynamic processes



$\Delta U = 0$ (an isothermal process)

There is no change in the internal energy of the gas. That is, its temperature is constant. Therefore $Q = 0 + W$; $Q = W$.

In an isothermal expansion (B → A) all the work done by the gas on the surroundings is supplied by the thermal energy transferred into the gas. In an isothermal compression (A → B), the work done on the gas is all transferred away from the gas as thermal energy. For a process to approximate to the ideal of being isothermal, the change must be as slow as possible. For an isothermal change:

$$pV = \text{constant (Boyle's law)}$$

$W = 0$ (an isovolumetric process)

There is no work done by or on the gas. This means there is no change in volume. Therefore $Q = \Delta U + 0$; $Q = \Delta U$.

In this straightforward process, if thermal energy is transferred into the gas it simply gains internal energy and its temperature rises. If thermal energy is transferred away from the gas its internal energy and temperature decrease.

$Q = 0$ (an adiabatic process)

No thermal energy is transferred between the gas and its surroundings. Therefore $0 = \Delta U + W$; $\Delta U = -W$ for a compression and $-\Delta U = W$ for an expansion.

In an adiabatic expansion (B → A) all the work done by the gas is transferred from the internal energy within the gas, ΔU is negative and the temperature decreases. In an adiabatic compression (A → B) all the work done on the gas ($-\Delta W$) is transferred to the internal energy of the gas, which gets hotter. When gas molecules hit the inwardly moving piston they gain kinetic energy. Note that adiabatic lines on *pV* diagrams must be steeper than isothermal lines simply because, in equal expansions, the temperature falls during an adiabatic change but not (by definition) during an isothermal change.

For a process to approximate to the ideal of being adiabatic, the change must be as rapid as possible in a well-insulated container.

For an adiabatic change of an *ideal monatomic gas*:

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

This equation is given in the *Physics data booklet*.

For gases other than monatomic gases, V is raised to a different power, but this is not included in this course. Comparing this to the equation for an isothermal change (pV = constant), we see that similar changes in volume are associated with bigger changes in pressure in adiabatic changes. This is because there are also accompanying temperature changes.

$\Delta p = 0$ (an isobaric change)

Any expansion or compression that occurs at constant pressure. Therefore $Q = \Delta U + W$.

Isobaric changes usually occur when gases are allowed to expand or contract freely when their temperature changes, keeping their pressure the same as the surrounding pressure.

Worked examples

10 An ideal gas of volume 0.08 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 What is the change in the internal energy of the gas?

a An isobaric process.

b $W = p\Delta V$

$= (1.4 \times 10^5) \times (0.11 - 0.08)$

$= 4200$ J

c $\Delta Q = \Delta U + \Delta W$

$(7.4 \times 10^3) = \Delta U + 4200$

$\Delta U = +3200$ J

11 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 is increased.

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

$$\frac{p_2}{p_1} = \left(\frac{V_1}{V_2}\right)^{\frac{5}{3}} = 8.0^{\frac{5}{3}}$$

$$\log \frac{p_2}{p_1} = \frac{5}{3} \log 8.0 = 1.505$$

$$\frac{p_2}{p_1} = 32$$

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.

32 What temperature rise is produced when 0.45 kJ is transferred to the internal energy of 0.50 mole of an ideal gas at constant volume?

33 Explain the differences between the concepts of internal energy, thermal energy and temperature.

34 When a fixed mass of a particular gas expands by a certain amount, compare the amount of work that would be done in isobaric, adiabatic and isothermal changes. Explain your answer.

35 When 100 J of thermal energy was transferred to a gas it expanded and did 30 J of work. What was the change in internal energy?

36 Explain why a gas gets hotter when it is compressed rapidly.

37 When a gas expanded isothermally, 3000 J of work were done.
a Was the work done *on* the gas or *by* the gas?
b Was the thermal energy transferred into the gas zero, 3000 J, higher than 3000 J or lower than 3000 J?

38 The piston of a bicycle pump is pulled out *very slowly*.
a What happens to the temperature, internal energy and pressure of the gas?
b What assumption did you make?

39 Explain why an adiabatic expansion shown on a *pV* diagram must be steeper than an isothermal expansion starting at the same point.

40 A sample of an ideal gas expands from 2.23 m^3 to 3.47 m^3. The gas pressure was originally 3.63×10^5 Pa.
a If the change was isothermal, what was the final pressure of the gas?
b Draw a suitable *pV* diagram and include these two states.
c Draw a line on the diagram to show how the gas went between these two states.
d Estimate the work done during expansion of the gas.

41 An ideal monatomic gas expands adiabatically from a volume of 6.78×10^{-2} m^3 to 9.53×10^{-2} m^3. If the initial pressure of the gas was 4.32×10^5 Pa, what was its final pressure?

Cyclic processes and pV diagrams

An expanding gas (sometimes called the **working substance**) can do useful work – for example by making a piston move along a cylinder. However, it cannot expand for ever, so any practical device transferring thermal energy to mechanical work must work in **cycles**, involving repeated **expansion** followed by **compression**, followed by expansion etc. In this section we will discuss some of the physics principles that are fundamental to cyclical processes in an ideal gas. However, it is important to realise 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 causing expansion while useful mechanical work is done by the gas. This is represented by the path AB in Figure 14.30a. The shaded area under the graph represents the work done *by* the gas in this process.

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 14.30b. The area under this graph represents the work done *on* the compressed gas *by* the surroundings. The difference in areas, shown in Figure 14.30c, 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 useful work done and no energy wasted.

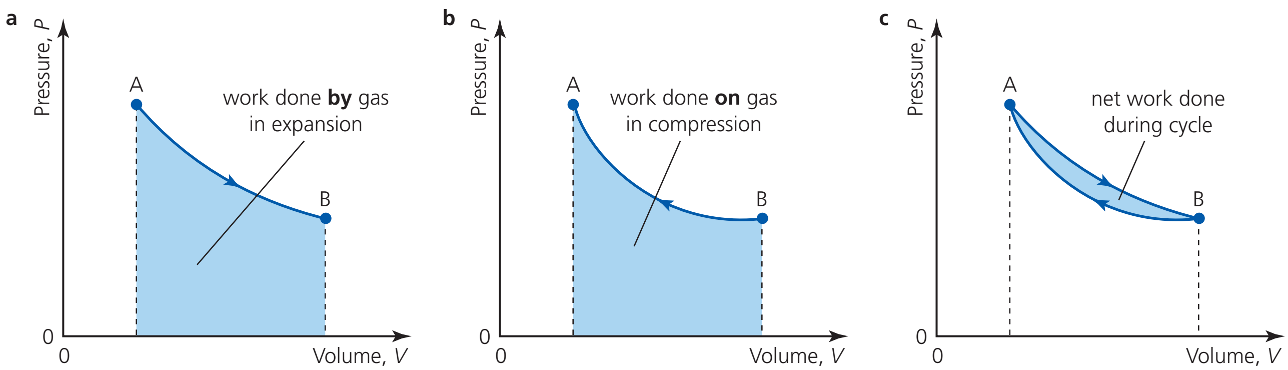


Figure 14.30 Work done in a thermodynamic cycle

In general, the efficiency, η, of any process (see Chapter 2) is defined by:

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

This equation is given in the *Physics data booklet*.

Worked examples

12 Determine the efficiency of the simple cycle shown in Figure 14.30 if the area shown in (a) was 130 J and the area in (b) was 89 J.

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

13 Figure 14.31 shows one simplified cycle, ABCA, of an ideal 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?

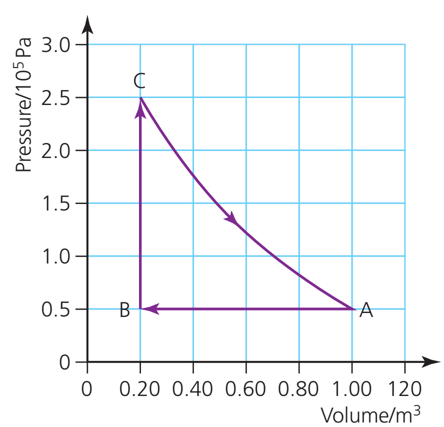


Figure 14.31

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

$$= \frac{1}{2} \times (0.85 - 0.20) \times (2.5 - 0.50) \times 10^5$$

(estimated from a triangle having about the same area, as judged by eye)

$W = 6.5 \times 10^4$ J

d efficiency = useful work output/total energy input

$$\frac{6.5 \times 10^4}{1.3 \times 10^5} = 0.50 \text{ (or 50\%)}$$

Carnot cycle

The thermodynamic cycle that produces the maximum theoretical efficiency for an ideal gas is shown in Figure 14.32 – it is called the *Carnot cycle*.

It is a four-stage process: 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 in adiabatic changes.)

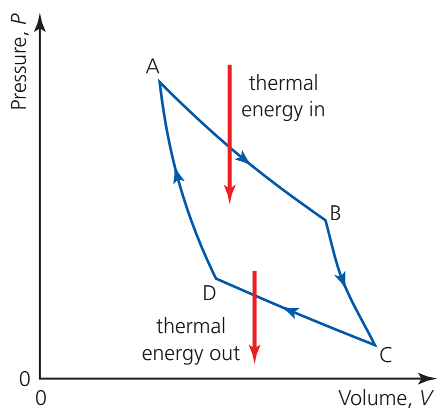


Figure 14.32 The most efficient thermodynamic cycle

Utilizations

Diesel engines

In car engines (internal combustion) a fuel and air mixture is ignited to cause the rapid, controlled explosions that provide power to the vehicle. In cars with petrol (gasoline) engines, this is achieved using an electrically produced spark across a 'spark plug' inserted into the gas cylinder. Diesel engines work on a different principle.

Changes to a gas that occur rapidly are approximately adiabatic, so that the work done on the gas when it is compressed causes a rapid increase in temperature. If the compression ratio is high enough (maybe 20 : 1), the rise in temperature can be large enough to ignite the gas. (The compression ratio is the ratio maximum volume/minimum volume of the working space of the cylinder, as the piston moves up and down.) This is what happens in a diesel engine, a type of engine named after the German Rudolf Diesel, who produced the first successful engine without spark plugs in 1897.

Because of their higher compression ratios, diesel engines need to be stronger than other engines and they also use a different fuel. They are more efficient than other forms of internal

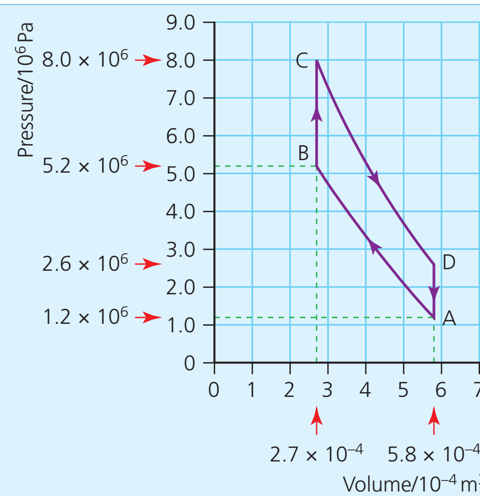
■ **Figure 14.33** Diesel-powered train

combustion engines and they are widely used where large and powerful engines are needed, especially when the extra mass of a diesel engine is not an important factor, such as in trains (see Figure 14.33) and ships. They have become more and more popular for cars in recent years and about half the cars sold in Europe are now powered by diesel engines.

1 Suppose you were about to buy a new car. Carry out research and a make a list of the advantages and disadvantages of owning a diesel-powered car (compared with a petrol-powered car). Which would you choose?

42 Figure 14.34 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?

43 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 when work is done *on* the gas.
- **f** Estimate the net work done *by* the gas during the cycle.

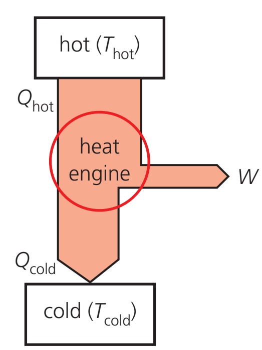


■ **Figure 14.34**

■ Thermal efficiency

The flow of energy during the continuous conversion of thermal energy to mechanical work in a heat engine is represented in Figure 14.35. A temperature difference, $T_{hot} - T_{cold}$, is needed between hot and cold reservoirs so that there is a resulting flow of thermal energy, which operates the engine. Thermal energy Q_{hot} flows out of the hot reservoir and Q_{cold} flows into the cold reservoir. The difference in thermal energy is transferred to doing useful mechanical work, W.

$$\text{thermal efficiency, } \eta = \frac{\text{useful work done}}{\text{energy input}} = \frac{W}{Q_{hot}}$$



■ **Figure 14.35** Energy flow in a heat engine

It is not possible to convert all of the thermal energy into work. (This is a version of the second law of thermodynamics, which is introduced later in this chapter.) For more than 200 years scientists and engineers have spent considerable amounts of time and effort in trying to improve the efficiency of heat engines of various designs. Unfortunately, the laws of physics limit the efficiency that can be achieved. The following simple equation can be used to calculate the maximum efficiency possible with a Carnot cycle.

$$\eta_{carnot} = 1 - \frac{T_{cold}}{T_{hot}}$$

This equation is given in the *Physics data booklet*.

Worked example

14 a Determine the maximum theoretical thermal efficiency of a heat engine operating with an inlet temperature of 300°C and an outlet temperature of 150°C.

b What would the thermal efficiency become if both temperatures were reduced by 50°C?

a $\eta_{\text{carnot}} = 1 - \frac{T_{\text{cold}}}{T_{\text{hot}}} = 1 - \frac{423}{573} = 0.26$ or 26%

b $\eta_{\text{carnot}} = 1 - \frac{T_{\text{cold}}}{T_{\text{hot}}} = 1 - \frac{373}{523} = 0.29$ or 29%

A **heat pump** works like a heat engine but in reverse, using a work input to enable the transfer of thermal energy from colder to hotter, as in a refrigerator or an air-conditioner. It is possible to remove thermal energy from the environment and use it to help in keeping a house warm in winter, although in many circumstances this may not be practicable.

44 a Calculate the maximum theoretical efficiency of a heat engine operating at an input temperature of 200°C and releasing heat to the surroundings at 40°C.

b How could engineers try to improve that efficiency?

c Can you think of any way in which this wasted energy could be used?

45 Draw a schematic diagram, similar to Figure 14.35, to represent the principle of operation of a heat pump.

46 Use the internet to investigate the circumstances under which it may be efficient to use a heat pump to warm a house.

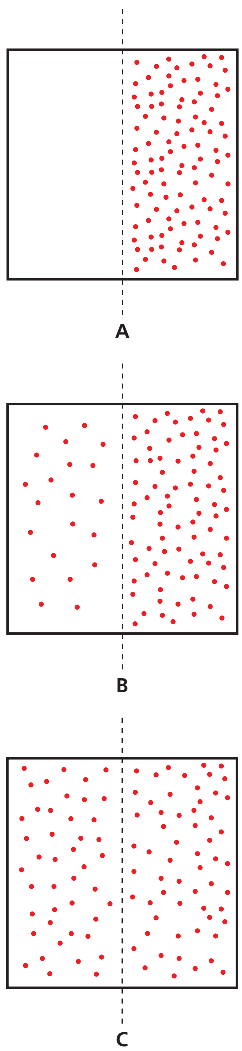


Figure 14.36 Gas molecules spreading out in a container

Order and disorder

The molecules of an ideal gas move in a completely random and uncontrollable way. What happens to them is simply the most likely (probable) 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 the three diagrams in Figure 14.36, which show the distribution 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 (maybe 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 the same way, statistically, we would not believe that if 100 coins were tossed, they could all land 'heads' up.) These diagrams only show about 100 molecules drawn to represent the 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.

Statistical analysis shows that we can be very sure that *every* isolated system of any kind will become more disordered as time progresses. Put simply, this is because everything is made up of particles, and individual atoms and molecules are usually 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 molecular order from disorder requires intervention and is difficult, or even impossible. There are a countless number of ways to disorganize a system, but only few ways to organize it.

Entropy

The disorder of a molecular system can be calculated. It is known as the **entropy** of the system. However, in this course we are not concerned with a precise mathematical definition of entropy.

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

Molecular disorder and the concept of entropy are profound and very important ideas. It is 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

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.

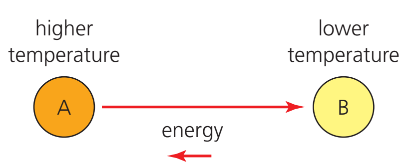
This is sometimes expressed by the statement 'entropy can never decrease'. But it should be stressed that it is certainly possible to reduce the 'local' entropy of *part* of a system, but in the process another part of the system will gain even more entropy. For example, the growth of a plant, animal 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.

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*.

Alternative ways of expressing the second law

Consider any two objects at different temperatures placed in thermal contact in the same system with no external influences, as shown in Figure 14.37.

Figure 14.37 Exchanges of energy



Energy can flow from A to B and from B to A, but the flow of energy from A to B is more likely because the energy is more ordered (concentrated) in A. Therefore, the *net* flow of energy will always be from hotter to colder. Obviously, this is a matter of common observation, but it is also the basis of an alternative version of the second law, first expressed by Rudolf Clausius:

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

But we can use *machines* to transfer energy from colder to hotter by doing external work (heat pumps – see p. 24).

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 (Kelvin–Planck form) of the second law of thermodynamics is expressed as:

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

Nature of Science

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 here 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.

Entropy changes

In all thermodynamic processes, total energy is always conserved and entropy always increases.

It is not possible to study physics for long without becoming familiar with the fact that in every macroscopic process some useful energy is 'lost' or 'wasted'.

As an example, consider a bouncing ball – the height of each bounce gets lower and lower as shown in Figure 14.38a. If you were to see the opposite happening, with a ball bouncing higher and higher, as in Figure 14.38b, you would probably be amazed, thinking it impossible unless the ball contained an unseen source of energy.

As the ball moves through the air and as it bounces on the ground, frictional and contact forces result in an increase of internal energy (the ball gets warmer) and then thermal energy is dissipated into the surroundings. We can now interpret this in terms of the second law of thermodynamics and entropy. Figure 14.38b shows an 'impossible' situation because it would involve a decrease in entropy of the system as the energy became more ordered.

Because the molecules in the ball are all moving in the same direction, we can describe the kinetic energy of the ball as *ordered* (disregarding the ball's initial internal energy). As the ball's temperature increases, the increased random kinetic energy of the molecules is *disordered* energy. The overall entropy of the universe increases as the ball loses kinetic energy and gains internal energy.

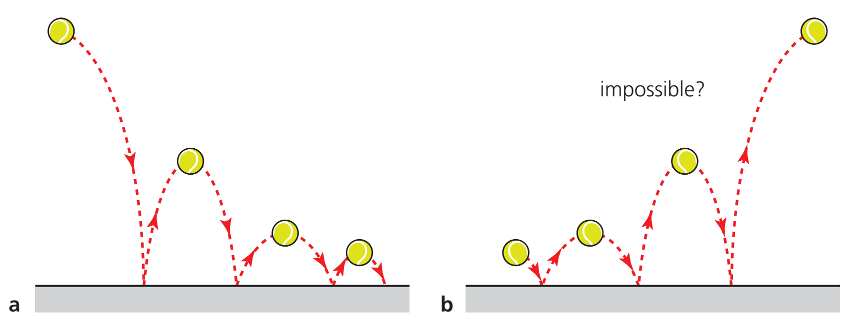


Figure 14.38 Bouncing ball

The same principles apply to all macroscopic processes. We can say that in such processes some, or all, of the energy is *degraded* because it has been transferred from a useful form to a form in which it can no longer be useful to us; that is, it can no longer perform useful work. The same ideas can be expressed in everyday terms in many different ways – for

■ **Figure 14.39** A refrigerator transfers thermal energy from the food and reduces entropy, but where does the energy go?

example, we may talk casually about 'energy being shared out, lost or wasted'. But a higher level of understanding is suggested by reference to thermal energy being *dispersed* or *dissipated* to the surroundings.

The dissipation and degradation of energy occurs because of the chaotic behaviour of molecules. It is simply much more likely that energy will be spread out and become disordered. Given any opportunity, that is exactly what the random behaviour of molecules will *always* produce. This profound and very important idea explains why we are surrounded by 'one-way' processes – events that simply could not happen in reverse. The concepts of **irreversibility**, the 'arrow of time', and why so many things cannot happen even though the principle of conservation of energy is not broken, can all be explained using the powerful ideas of entropy and the second law of thermodynamics.

It may seem that there are some exceptions to the second law. Refrigerators (and air-conditioners) are designed to make something that may already be cooler than the surroundings even cooler by transferring thermal energy from colder to hotter. But these are not 'spontaneous' energy transfers, and in the cooling processes the *overall* entropy of the refrigerator *and its surroundings* will have increased. The *local* entropy of the contents of the refrigerator will not decrease as much as the entropy of the surroundings will increase (Figure 14.39).

47 What happens to the internal energy and the entropy of an ideal gas when it undergoes an isothermal expansion?

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

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

50 Is it possible to make a video recording of a simple event that would *not* look ridiculous if shown in reverse? Explain.

51 There are four laws of thermodynamics – only the first and second are included in this book. They are often 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?

ToK Link

Quotes from Richard Feynman

Richard Feynman (1918–1988) received the Nobel Prize for physics in 1965. Although primarily involved in research, he was also a leading personality in the world of physics for many years. He is especially well known for his lectures, books and TV appearances in which he tried to explain the complexities of physics to the general public. The following quotations give a flavour of his thought-provoking approach to science.

'If, in some cataclysm, all scientific knowledge were to be destroyed, and only one sentence passed on to the next generation of creatures, what statement would contain the most information in the fewest words?

I believe it is the atomic hypothesis (or atomic fact, or whatever you wish to call it) that all things are made of atoms – little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another. In that one sentence you will see an enormous amount of information about the world, if just a little imagination and thinking are applied.'

'Science is the belief in the ignorance of experts.'

'You can know the name of a bird in all the languages of the world, but when you're finished, you'll know absolutely nothing whatever about the bird… So let's look at the bird and see what it's doing – that's what counts. I learned very early the difference between knowing the name of something and knowing something.'

'A great deal more is known than has been proved.'

'We can't define anything precisely. If we attempt to, we get into that paralysis of thought that comes to philosophers… one saying to the other: "you don't know what you are talking about!". The second one says: "what do you mean by talking? What do you mean by you? What do you mean by know?"'

'Our imagination is stretched to the utmost, not, as in fiction, to imagine things which are not really there, but just to comprehend those things which are there.'

Figure 14.40 Richard Feynman

'I have approximate answers, and possible beliefs, and different degrees of certainty about different things, but I'm not absolutely sure of anything… but I don't have to know an answer. I don't feel frightened by not knowing things, by being lost in a mysterious universe without having any purpose, which is the way it really is, as far as I can tell, possibly. It doesn't frighten me.'

The Pleasure of Finding Things Out: The Best Short Works of Richard Feynman, edited by Jeffery Robbins

'From a long view of the history of mankind – seen from, say, ten thousand years from now, there can be little doubt that the most significant event of the 19th century will be judged as Maxwell's discovery of the laws of electrodynamics. The American Civil War will pale into provincial insignificance in comparison with this important scientific event of the same decade.'

'The exception tests the rule. Or, put another way, the exception proves that the rule is wrong. That is the principle of science. If there is an exception to any rule, and if it can be proved by observation, that rule is wrong.'

1 You may detect in some of these quotes a little disrespect for accepted scientific procedures. Do you think that such an attitude is useful for a good scientist? Explain your answer.

Calculating changes in entropy

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:

$$\Delta S = \frac{\Delta Q}{T}$$

This equation is given in the *Physics data booklet*. The units of entropy change are $\mathrm{J\,K^{-1}}$.

The use of this equation will normally be restricted to examples in which the thermal energy transfer is to or from a system or surroundings that can be assumed to remain at constant temperature.

Worked example

15 Determine the increase in entropy of 500 g of ice when it melts at 0°C (specific latent heat of fusion of water, $L_f = 3.34 \times 10^5\,\mathrm{J\,kg^{-1}}$).

$$\Delta S = \frac{mL_f}{T} = \frac{(0.500 \times 3.34 \times 10^5)}{273}$$
$$= 612\,\mathrm{J\,K^{-1}}$$

52 During a chemical reaction 48 kJ were transferred as thermal energy into a laboratory. Estimate the increase in entropy of the laboratory.

53 **a** Determine the change in entropy of 100 g of ice cream when it freezes at 0 °C. ($L_f = 1.9 \times 10^5\,\mathrm{J\,kg^{-1}}$).
b Suggest a value for the change in entropy of the refrigerator and surroundings during this process.

54 Estimate the increase in entropy of a kitchen at a temperature of 18 °C when a 200 g cup of coffee cools from 80 °C to 60 °C.

14.3 (B3: Additional Higher) Fluids and fluid dynamics – *fluids cannot be modelled as point particles; their distinguishable response to compression from solids creates a set of characteristics that require an in-depth study*

In this section we will discuss the properties of substances that can flow: gases and liquids, which are collectively described as *fluids*. In everyday language it is common to describe a liquid as a 'fluid', and much of this chapter will be related to liquids rather than gases. In macroscopic terms the essential differences between a gas and a liquid are that liquids are constrained by a surface and cannot be compressed, whereas gases are relatively easy to compress and will diffuse to fill their containers.

In the first part of the work on fluids we will discuss the properties of *static* (unchanging, not moving) fluids – this area of study is commonly called **hydrostatics**. Then we will look at fluids in motion – *fluid dynamics*.

Density and pressure

From earlier work, we know that

- density = mass/volume, $\rho = m/V$; units $\mathrm{kg\,m^{-3}}$. The density of pure water is $1.0 \times 10^3\,\mathrm{kg\,m^{-3}}$, the density of the atmosphere at sea level is $1.2\,\mathrm{kg\,m^{-3}}$, both at 20°C.
- pressure = force/area; $p = F/A$; units $\mathrm{N\,m^{-2}}$, also known as pascals, Pa. The **pressure of the atmosphere** at sea level is 1.0×10^5 Pa.

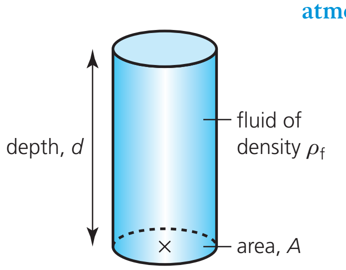


Figure 14.41 Pressure underneath a fluid at point X

The **hydrostatic pressure** under a fluid due to the weight of the fluid pressing downwards can be calculated by reference to Figure 14.41, which shows a cylinder of fluid of depth d, area A and density ρ_f.

$$\text{pressure at X due to the fluid, } p = \frac{F}{A} = \frac{\text{weight of fluid}}{\text{area of base}} = \text{volume} \times \text{density} \times \frac{g}{A}$$

$$= \frac{Ad\rho_f g}{A}$$

$$p = \rho_f g d$$

This equation can be applied to any static fluid in any shaped container, or no container. The equation shows that, for a given fluid, the pressure depends *only* on depth (assuming we are referring only to locations where g is constant). This is demonstrated in Figure 14.42, in which the liquid depth in each of the four tubes is the same, and so produces the same pressure at the bottom of the container. If this were not true there would be pressure differences causing movement of the liquid.

The pressure under a solid object (such as a book resting on a table) acts only downwards in the same direction as its weight, but pressure in fluids is caused by the random motion of particles, so that:

pressure due to fluids acts equally in all directions.

Figure 14.42 Pascal's vases

In Figure 14.43 air pressure acting *upwards* is enough to stop the water falling out of the glass because of the water pressure acting downwards.

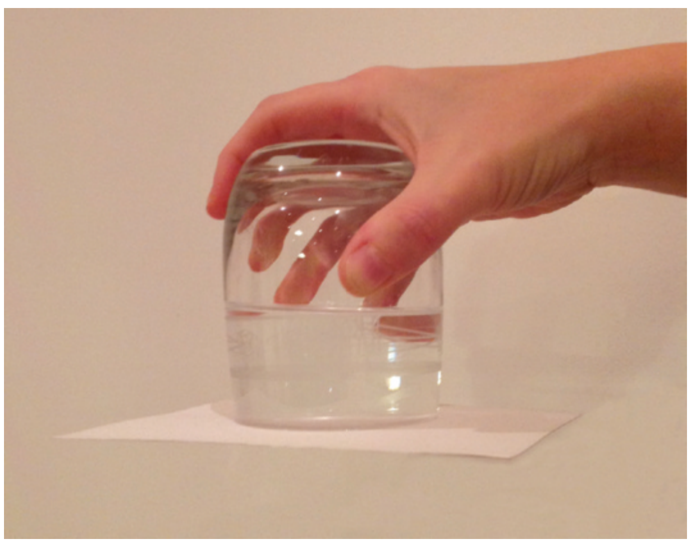


Figure 14.43 Air pressure supporting water in a glass

Worked example

16 Estimate a value for the pressure underneath the Earth's atmosphere assuming that the height of the atmosphere is 10 km and the average density of air is 1.0 kg m^{-3}.

$$p = \rho_f g d = 1.0 \times 9.81 \times 10^4 \approx 10^5 \text{ Pa}$$

If one fluid is on top of another, the total pressure is the sum of the individual pressures.

The equation $p = \rho_f g d$ is most commonly applied to liquids and, if we want to know the pressure, p at a certain level in a liquid that is exposed to the atmosphere, we must add the pressure due to the liquid to the pressure due to the atmosphere, p_0, above it, so that the equation becomes:

$$p = p_0 + \rho_f g d$$

This equation is given in the *Physics data booklet*.

Worked example

17 How deep below the surface of a freshwater lake would you need to go before the total pressure was 2.84×10^5 Pa? (Normal atmospheric pressure = 1.01×10^5 Pa; density of water in lake = 998 kg m^{-3}.)

$p = p_0 + \rho_f g d$

$2.84 \times 10^5 = (1.01 \times 10^5) + (998 \times 9.81 \times d)$

$d = 18.7$ m

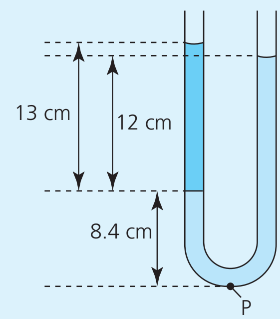
55 What depth of mercury, which has a density of 1.35×10^4 kg m^{-3}, will produce the same pressure as the atmosphere (1.01×10^5 Pa)?

56 Explain why the tubes in Figure 14.42 must be open at the top.

57 How far under seawater of density 1.03×10^3 kg m^{-3} would divers need to be for the pressure on them to be four times the atmospheric pressure on the surface?

58 Figure 14.44 shows a U-tube containing pure water of density 1.00×10^3 kg m^{-3} and olive oil (which has a lower density).
- **a** Which side has the olive oil?
- **b** Determine the density of the olive oil.
- **c** Calculate the total pressure at point P (assume atmospheric pressure = 1.013×10^5 Pa).

59 What is the total pressure 58.3 m below the Dead Sea (Figure 14.45), which itself is 427 m below normal sea level (assume atmospheric pressure at normal sea levels is 1.01×10^5 Pa, density of Dead Sea water is 1240 kg m^{-3}, density of air at 30 °C is 1.16 kg m^{-3})?



■ **Figure 14.44**

■ **Figure 14.45** Floating in the Dead Sea

■ Buoyancy and Archimedes's principle

Buoyancy is the ability of a fluid to provide a vertical upwards force on an object placed in or on the fluid. This force is sometimes called **upthrust** and it can be explained by considering the pressures on the upper and lower surfaces of the object. Figure 14.46 shows a cylinder of cross-sectional area A and depth d immersed in a fluid of density ρ_f.

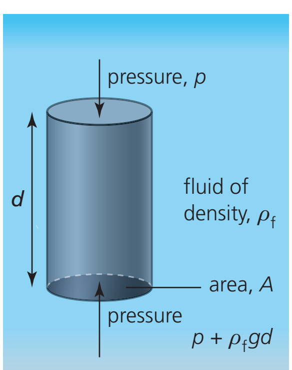
Because of the increased depth, the pressure on the lower surface will be greater than the pressure on the upper surface by an amount $\rho_f g d$. Therefore, there is a buoyancy force, B, acting upwards given by:

$$B = \text{extra pressure} \times \text{area} = \rho_f g d \times A$$

Or, because the volume of the object = volume of the fluid displaced, $V_f = dA$,

$$B = \rho_f V_f g$$

This equation is given in the *Physics data booklet*.



■ **Figure 14.46** Cylinder immersed in a fluid

Although this equation has been derived for a cylindrical solid immersed in a fluid, it can also be applied to other shapes. It also applies to objects *floating* on liquids, in which case V_f = volume of the fluid displaced, *not* the total volume of the floating object.

Since $\rho_f V_f$ is equal to the mass of the fluid displaced, $\rho_f V_f g$ is equal to the weight of the fluid displaced. The equation above can then be expressed as follows:

> When an object is wholly or partially immersed in a fluid, it experiences an upthrust (buoyancy force) equal to the weight of the fluid displaced.

This classic physics principle was first stated by Archimedes more than 2250 years ago. A range of different experiments can be designed to investigate it.

The object shown in Figure 14.47 will rise if $B > mg$ and it will fall (sink) if $B < mg$ (as shown). If a solid object is a made of only one material, it will rise if its density is less than the surrounding fluid and fall if its density is higher than the surrounding fluid.

The boat shown in Figure 14.48 is floating. Because it is floating, we know that the weight of the fluid displaced (represented by the shaded area) must be equal to the weight of the object. If more weight is added in or on the boat, it will sink lower into the fluid until it displaces a weight of fluid equal to the new weight of the object. If this is not possible, the object will sink.

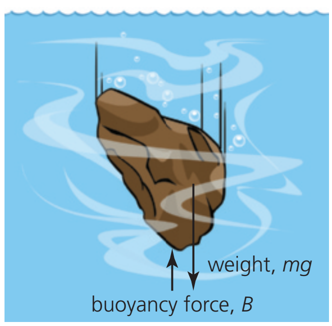


Figure 14.47 Forces on an object immersed in a fluid

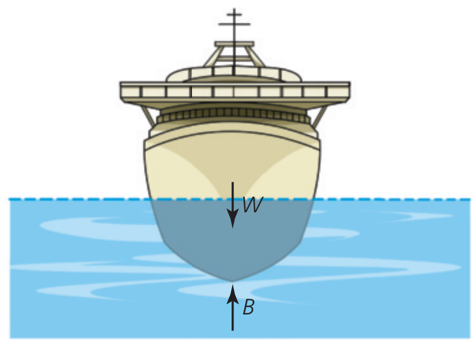


Figure 14.48 A floating object

60 Explain how a heavy oil tanker of mass 500 000 tonnes is able to float on water.

61 a A wooden cube with a density of $880\,\text{kg}\,\text{m}^{-3}$ is floating on water (density $1000\,\text{kg}\,\text{m}^{-3}$). If the sides of the cube are 5.5 cm long and the cube is floating with a surface parallel to the water's surface, what depth of wood is below the surface?

b It is commonly said that about 90 per cent of an iceberg is below the surface (Figure 14.49). Use this figure to estimate a value for the density of sea ice. Assume the density of sea water is $1025\,\text{kg}\,\text{m}^{-3}$.

62 Estimate the buoyancy force on yourself when you are immersed in (a) water and (b) air.

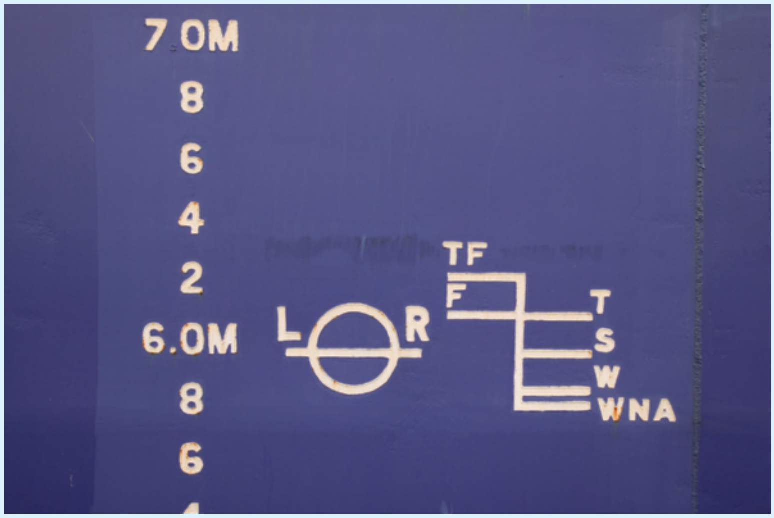
63 Learning to scuba dive involves being able to remain 'neutrally buoyant', so that the diver stays at the same level under water. Explain why breathing in and out affects buoyancy.

64 Find out how a submarine is able to move upwards and downwards in the ocean.

65 A balloon containing helium of total volume $6500\,\text{cm}^3$ is moving through the air. Will it rise or fall if its total mass is 5.4 g? (Density of air = $1.3\,\text{kg}\,\text{m}^{-3}$)

Figure 14.49

66 Figure 14.50 shows markings on the side of an ocean-going ship. Find out what they are called and what they represent.



■ **Figure 14.50**

ToK Link

Fact or fiction: the value of mythology in science

The mythology behind the anecdote of Archimedes's 'Eureka!' moment of discovery (of the buoyancy principle) demonstrates one of the many ways scientific knowledge has been transmitted throughout the ages. What role can mythology and anecdotes play in passing on scientific knowledge? What role might they play in passing on scientific knowledge within indigenous knowledge systems?

Science is often represented as an unemotional and inflexible human activity. Personalising the history of science with interesting or eccentric individuals with whom people can identify is useful in combating this impression. Iconic moments, such as Archimedes and his bath or an apple falling on Newton, are easily remembered and visualized, and in this respect it matters little whether they are true or myths.

■ Pascal's principle

Because any liquid is **incompressible** (its volume cannot be reduced) and its molecular motions are random, we can state the following principle:

A pressure exerted anywhere in an enclosed static liquid will be transferred equally to all other parts of the liquid.

If different parts of the liquid are at different heights, however, this will result in additional differences in pressure, which may or may not be significant. The most important application of this principle is in *hydraulic machinery*.

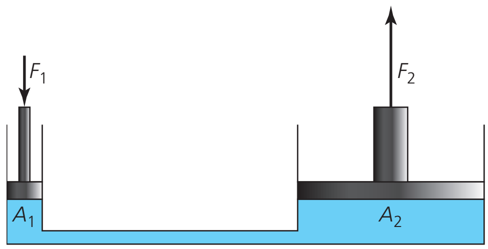
Hydraulic machinery

In hydraulic machines a force is applied to an incompressible liquid (usually a kind of oil) creating a pressure that is transferred using tubes (which may be flexible) to another location where a *larger* force is created. See Figure 14.51, which shows a sketch of simple apparatus that can be used to investigate this effect.

A force, F_1, acting on a piston of area A_1 on the cylinder on the left-hand side, creates a pressure that is transferred to the right-hand side, creating a force, F_2, acting on a piston of area A_2.

Because the pressure is equal everywhere (Pascal's principle):

$$p = \frac{F_1}{A_1} = \frac{F_2}{A_2}$$



■ **Figure 14.51** The principle of hydraulic machinery

So, in principle, any force can be changed into any other larger (or smaller) force by suitable choice of the areas. Multiplying a force using a machine does not break any laws of physics, but we cannot multiply energy! This means that the work done by the input force must be equal to the work done by the output force, assuming that the process is 100 per cent efficient. This is similar in principle to other 'force multiplying' machines such as simple levers, pulleys, jacks and ramps. If the machine is 100 per cent efficient (which, of course, is idealized):

$$F_1 \times \text{distance moved by } F_1 = F_2 \times \text{distance moved by } F_2$$

Hydraulic machines are in widespread use, especially wherever very large forces are required, such as in excavators of various designs. See Figure 14.52. The principle is also used in vehicle braking systems.

■ **Figure 14.52** Hydraulic machinery is used in this excavator

67 Consider Figure 14.51. Suppose $F_1 = 28\,\text{N}$, $A_1 = 2.1\,\text{cm}^2$ and $A_2 = 19.7\,\text{cm}^2$.

- **a** Calculate F_2.
- **b** If F_1 moves down 10 cm, how far does F_2 move up?
- **c** What value of F_1 would be needed to raise a load of 9.7 kg on the right-hand side if the machine was 92 per cent efficient?

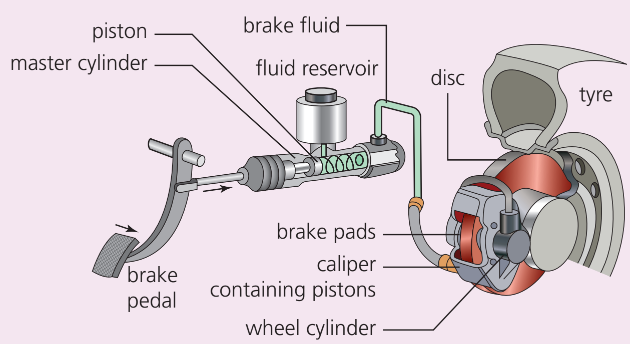
68 Summarize the advantages of using hydraulic machinery in, for example, the construction of a railway tunnel.

Utilizations

Hydraulic braking systems

Cars and other road vehicles are stopped because of friction between the road and the tyres. This occurs when forces are applied to stop the wheels rotating. These forces originate when the driver puts a foot on the brake pedal.

The force provided by the foot acts on a piston in a master cylinder to push the brake fluid (oil) along very strong flexible tubes all the way to the braking mechanisms close to the wheels. Figure 14.53 shows the connection to only one of the wheels. In the braking mechanisms at each wheel the fluid pressure acts on pistons to force the brake pads onto a metal disc on the wheel,



■ **Figure 14.53** Car braking system (only one wheel shown) creating fiction which slows its rotational speed. Because the total area of all the brake cylinders is much larger than the area of the master cylinder, the magnitude of the force is greatly increased.

1 a What average force is required to stop a 1000 kg car travelling at $10\,\mathrm{m\,s^{-1}}$ in 15 m?

b To achieve this braking force, what average force needs to be exerted on the master cylinder if it has an area of $1.0\,\mathrm{cm^2}$ and the four brake cylinders each have an area of $18\,\mathrm{cm^2}$?

2 Explain why it can be dangerous if there is any air in the braking system of a car.

Fluid dynamics

Hydrostatic equilibrium

We are familiar with the concept of equilibrium from other branches of physics, for example in the study of mechanics and thermal physics. 'Equilibrium' means that, although a system may be subjected to different influences, they are balanced, so the state of the system is not changing. Hydrostatic equilibrium may be seen as being similar to translational equilibrium in mechanics.

A fluid is in hydrostatic equilibrium if it is either at rest, or if any parts of it that are moving have a constant velocity.

This will occur when forces are balanced by differences in pressure. For example:

- A floating boat will be in a state of hydrostatic equilibrium (as shown in Figure 14.48) if its weight is balanced by pressure differences in the water.
- The Earth's atmosphere (as a whole) is in hydrostatic equilibrium because pressure differences across the atmosphere are balancing the effects of gravity on the air.
- Most stars are in hydrostatic equilibrium because the inwards gravitational attraction between the particles is opposed by the outwards pressure of the hot gases and radiation.

The ideal fluid

Nature of Science

Modelling a fluid

Real fluids differ from each other in the way that they flow because there are different forces within the fluids and with any containers in which they may be placed. Earlier in this option we discussed *ideal gases* and it is now appropriate to discuss a model of *ideal fluids*. However, in our discussion of fluid dynamics we will be more concerned about the macroscopic properties of fluids rather than any kinetic theory of the particles they contain. The properties of fluids cannot be fully explained by only the behaviour of the individual particles that they contain.

We can idealize and simplify the flow of a fluid as the movement of *layers* sliding over each other (like playing cards sliding over each other), without any movement of fluid *between* those layers. This is described as **laminar flow**. See Figure 14.54.

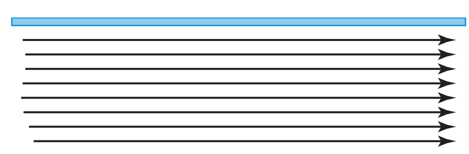


Figure 14.54 Idealized laminar flow of a fluid through a tube

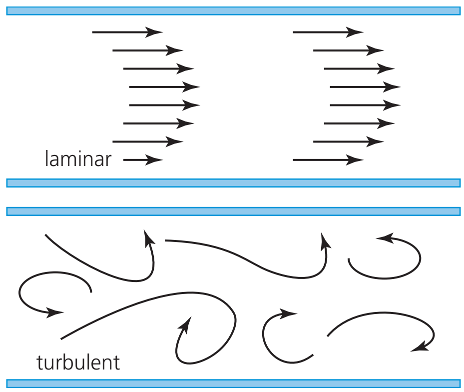
An ideal fluid:

- has constant density and is *incompressible*
- has constant pressure, acting equally in all directions
- is *non-viscous*: it has no frictional forces opposing motion. (It has zero *viscosity* – see later in this section). There are no shear forces between layers, or frictional forces between layers or any surfaces with which they may come in contact. (*Shear forces* are non-aligned parallel forces that tend to push a substance in opposite directions.)
- has a *steady flow* pattern that does not change with time and which can be represented by *streamlines*.

Streamlines

Streamlines are lines that show the paths that (mass-less) objects would follow if they were placed in the flow of a fluid.

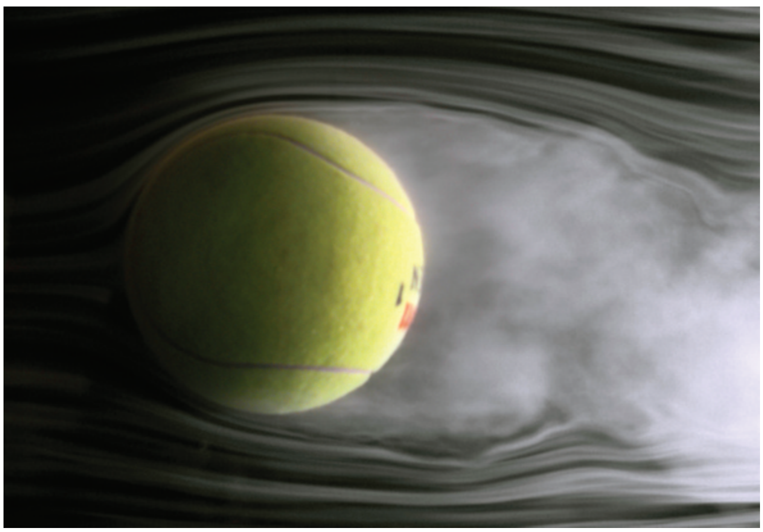
Figure 14.55 uses streamlines to represent the difference between laminar flow (also called **streamlined flow**)and non-laminar flow (*turbulence*):



■ **Figure 14.55** Using streamlines to represent laminar flow (with some viscosity) and turbulence in a pipe

- A tangent to a streamline shows the velocity vector of flow at that point.
- Streamlines cannot cross over each other.
- If the streamlines get closer together the fluid must be flowing faster.

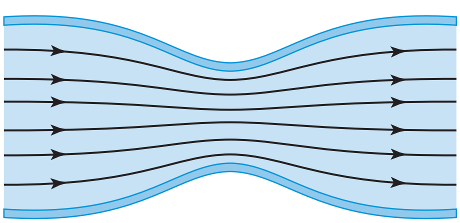
The flow of fluids around a moving object is often investigated in laboratories by keeping the object in a fixed position and making the fluid flow past it in a **wind tunnel** (or a liquid flow equivalent). Smoke or dye can be used to mark the streamlines. See Figure 14.56.



■ **Figure 14.56** Streamlines around a tennis ball in a wind tunnel

The continuity equation

Figure 14.57 represents the streamlines of an ideal fluid in a simple flow situation where a pipe gets narrower. The mass per second entering and leaving the tube must be constant and, because the fluid is incompressible, the flow speed must increase where the tube's cross-sectional area decreases.



■ **Figure 14.57**

The volume passing any point every second must be constant – it is equal to the cross-sectional area, A, multiplied by the speed of the fluid, v. That is, for the flow of an ideal fluid through any given enclosed system, at any point:

$$Av = \text{constant}$$

This is known as the *continuity equation* and it is given in the *Physics data booklet*. The constant is called the **(volume) flow rate**.

Worked example

18 Oil is flowing through a pipe at a volume flow rate of $0.014\,m^3\,s^{-1}$.

a What is the speed of the oil if the pipe has a radius of 12 cm?

b What is the speed of the oil at a point where the pipe narrows to a radius of 9.5 cm?

a $\pi r^2 \times v = 0.014$

$$v = \frac{0.014}{\pi\, 0.12^2} = 0.31\,m\,s^{-1}$$

b $\pi r^2 \times v = 0.014$

$$v = \frac{0.014}{\pi\, 0.095^2} = 0.49\,m\,s^{-1}$$

69 Water flowing at an average speed of $86\,cm\,s^{-1}$ in a river of width 5.32 m reaches a place where the river narrows to 2.76 m, but the depth of the river is unchanged at 4.1m.

a Estimate the speed of the water in the narrow section.

b If the speed of the water then decreases to $150\,cm\,s^{-1}$,while the width remains the same, what is the depth of the river?

70 Natural gas flows through a cylindrical pipeline of radius 48 cm at a speed of $3.9\,m\,s^{-1}$. It is kept moving by the pressure provided by a pump with an effective diameter of 26 cm.

a What is the maximum speed of the gas through the pump?

b State one assumption that you made in answering (a).

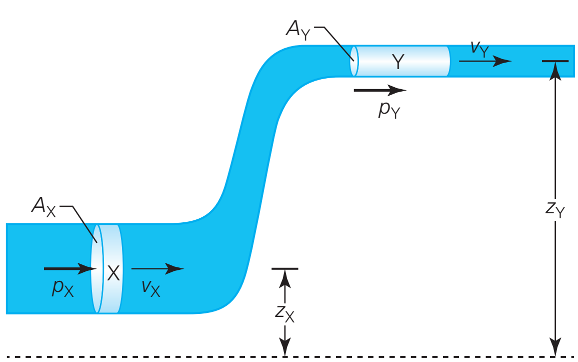
The Bernoulli equation and the Bernoulli effect

In general, we would expect that the speed of flow of an incompressible fluid in an enclosed system would increase if:

- some kind of pump was providing a pressure difference
- the pipe was going down to a lower level
- the pipe was getting narrower.

The *Bernoulli equation* involves these factors in an equation that describes the steady flow of an ideal fluid in any system.

Consider Figure 14.58 in which a fluid is moving through a tube from left to right. In this example the fluid moves to a narrower tube ($A_Y < A_X$) at a greater height, as measured from some arbitrary level ($z_Y > z_X$). The small volume, V (mass m and density ρ), of fluid shown at position X moves later to position Y.



■ **Figure 14.58** A fluid moving to a greater height in a narrower tube

When an ideal fluid flows through a system at a steady rate, we know that the fluid in the system can have different potential and kinetic energies, but the total energy must be constant. Using the principle of conservation of energy:

(kinetic energy of V + gravitational potential energy of V) at Y − (kinetic energy of V + gravitational potential energy of V) at X = work done on V by any difference in pressures

But work done, $W = p\Delta V$ (see earlier in this chapter), or $W = \Delta pV = (p_X - p_Y)V$, so that:

(kinetic energy of V + gravitational potential energy of V) at X + $p_X V$ = (kinetic energy of V + gravitational potential energy of V) at Y + $p_Y V$

Or:

$$\frac{1}{2}mv^2 + mgz + pV = \text{constant}$$

Dividing throughout by V we get:

$$\frac{1}{2}\rho v^2 + \rho gz + p = \text{constant}$$

This equation (the *Bernoulli equation*) is given in the *Physics data booklet*.

- For fluid flow in horizontal directions z will be constant, so that $\frac{1}{2}\rho v^2 + p = \text{constant}$.
- For fluid flowing freely out of a container, the pressure at the top and at the open hole will be (almost) equal at atmospheric pressure, so that $\frac{1}{2}\rho v^2 + \rho gz = \text{constant}$.

Worked example

19 Consider Figure 14.58. A small pump is moving water to a level 43 cm higher.

a If the water is flowing at 72.0 cm s^{-1} at the lower level, what pressure difference is required to produce a flow of 54 cm s^{-1} at the higher level?

b If the diameter of the lower pipe is 2.4 cm, what is the radius of the upper pipe?

a $(\frac{1}{2}\rho v^2 + \rho gz)_X + p_X = (\frac{1}{2}\rho v^2 + \rho gz)_Y + p_Y$

Taking the lower level to be at a height $z_X = 0$

$(\frac{1}{2} \times (1.0 \times 10^3) \times 0.720^2) + (1.0 \times 10^3 \times 9.81 \times 0) + p_X = (\frac{1}{2} \times (1.0 \times 10^3) \times 0.54^2) + (1.0 \times 10^3 \times 9.81 \times 0.43) + p_Y$

$259 + 0 + p_X = 146 + 4220 + p_Y$

pressure difference, $p_X - p_Y = 4100\,\text{Pa}$

b Using the continuity equation, $Av = \text{constant}$:

$0.012^2\pi \times 0.72 = \pi r^2 \times 0.54$

$r = 1.4\,\text{cm}$

Two applications of the Bernoulli equation

Liquid flowing out of a container

Figure 14.59 shows a container of water with three holes. At each of the holes the pressure is equal to the atmospheric pressure, which means that there is no significant pressure difference between the top surface and the holes. Assuming that the water has insignificant kinetic energy at the surface, Bernoulli's equation reduces to:

$$\frac{1}{2}\rho v^2 = \rho gz$$

Where v is the horizontal speed of the emerging water and z is the depth of a hole beneath the surface. Dividing by ρ and rearranging we get:

$$v = \sqrt{2gz}$$

This is similar to an equation for the speed reached by a falling mass (equations of motion, Chapter 2). The parabolic trajectories of the water in Figure 14.59 clearly show how the velocity of the outlet increases with depth below the surface.

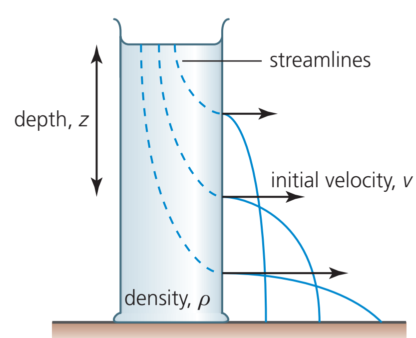


Figure 14.59 The 'leaky can' experiment

Pitot tubes

Pitot tubes are used to measure the speed of fluid flow (for example, through a pipe) or the speed of an object moving through a fluid. There are many different designs and applications. Figure 14.60 shows Pitot tubes mounted on an aircraft which are used to determine the speed of the plane relative to the air (its *air speed*), rather than speed relative to the land below (its *ground speed*).

Figure 14.60 Pitot tubes on an aircraft

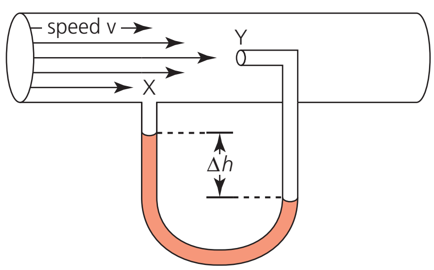


Figure 14.61 Simplified Pitot tube

The principle of a Pitot tube relies on comparing the pressure in the direct flow of the fluid to somewhere else *not* in the direct flow. Figure 14.61 represents as simplified example. (The different pressures could be measured by different devices along the same streamline, and the results compared electronically.)

Under conditions of hydrostatic equilibrium we can use Bernoulli's equation:

$$\frac{1}{2}\rho v_X^2 + \rho g z + p_X = \frac{1}{2}\rho v_Y^2 + \rho g z + p_Y$$

But, assuming the tube is horizontal, z is constant, so:

$$\frac{1}{2}\rho v_X^2 - \frac{1}{2}\rho v_Y^2 = p_Y - p_X = \Delta p$$

If we assume that the fluid colliding with the Pitot tube at Y has its velocity reduced to zero then:

$$\frac{1}{2}\rho v_X^2 = \Delta p$$

In the example shown in Figure 14.61 the pressure difference, Δp, can be determined from the U-tube manometer. This will be approximately equal to $\rho g \Delta h$ if the density of the manometer fluid is much higher than the density of the fluid flowing in the pipe and the rest of the Pitot tube.

The Bernoulli effect

For a fluid flowing horizontally (or with insignificant height variations), the Bernoulli equation reduces to:

$$\frac{1}{2}\rho v^2 + p = \text{constant}$$

This shows us that if an (ideal) fluid is flowing horizontally, or an object is moving horizontally through a fluid without turbulence, there must be a decrease in pressure wherever the speed increases.

This is commonly known as the *Bernoulli effect* and it has many interesting applications, including these three examples.

- *The curved path of spinning balls* – Figure 14.62 represents a spinning ball moving to the left through the air. The motion of the ball's surface will increase the speed of the air flow on one side (lower as shown) and decrease it on the other. This effect will be greater if the surface of the ball is *not* smooth. The difference in air speeds causes a pressure difference and a force in the direction shown.
- *Aircraft wings* – the cross-sectional shape of an aircraft wing (called its **aerofoil** or *airfoil*) will affect the way in which the air flows past it. If the shape causes the streamlines to be closer together above the wing, this increases the speed of the air and reduces the pressure, causing an upwards force called *lift*. See Figure 14.63. The effect may be increased by raising the leading edge of the aerofoil; this also causes the force of the air striking the aerofoil to have a vertical component, increasing lift.
- **Venturi tubes** – a fluid flowing through a tube will have less pressure at a place where the tube is narrower because the fluid must flow faster. See Figure 14.64.

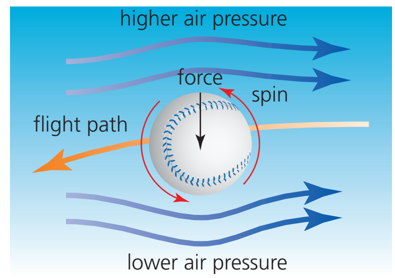


Figure 14.62 A ball can be made to curve in flight by making it spin

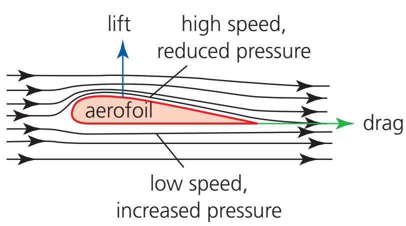


Figure 14.63 Action of an aerofoil

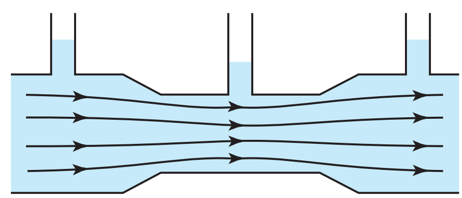


Figure 14.64 The Venturi effect

Measuring the decrease in pressure caused by a Venturi tube can be used to determine a fluid's flow rate.

Worked example

20 a Consider Figure 14.64. If the speed of the fluid entering through a tube of cross-sectional area 42 cm^2 on the left is 1.2 m s^{-1}, what is the speed of the fluid in the central section, where the cross-sectional area is 28 cm^2?

b If the density of the fluid is 870 kg m^{-3}, determine the difference in fluid pressure between the central section and the rest of the tube.

a Av = constant

$42 \times 1.2 = 28v$

$v = 1.8\ \text{m s}^{-1}$

b $\frac{1}{2}\rho v^2 + \rho g z + p = \text{constant}$

But z is constant, so:

$\frac{1}{2}\rho {v_X}^2 - \frac{1}{2}\rho {v_Y}^2 = p_Y - p_X = \Delta p$

$\Delta p = \frac{1}{2} \times 870 \times 1.2^2 - \frac{1}{2} \times 870 \times 1.8^2$

$= -780\ \text{Pa}$

The Venturi effect can be very useful in situations where fluids need to be mixed. A narrowing in a tube with one fluid flowing though it can produce a decrease in pressure that encourages another fluid to flow into the tube. For example, this is used in car engines to mix air and gasoline (petrol).

Nature of Science

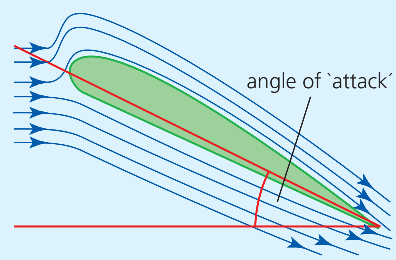
Fluids flowing across international borders

The storage of water and the control of its flow has been a necessary and long-established technology in many parts of the world for thousands of years. Water is such an important resource that, in places where it is scarce, it has been the cause of international disputes between countries. Dams and irrigation systems are seen as a benefit by some people and an ill-judged interference with the environment by others. With the world's population rising significantly, especially in some of the driest regions, and the effects of global warming uncertain, such issues may become more common.

The flow of gas and oil in pipelines over long distances and across international borders can be threatened at times of economic and political instability.

71 Figure 14.65 shows an aircraft wing that is moving steeply upwards. The air striking the lower surface causes a force perpendicular to the surface.

- **a** Copy the aerofoil and add a vector to represent the resultant force from the air.
- **b** Show the lift and drag components of the resultant force from the air.
- **c** Use information from the internet to discuss whether the lift force from the air striking the aerofoil, or the lift from the Bernoulli effect is greater.
- **d** Investigate the circumstances under which it can be possible for an aircraft to fly upside down.



■ **Figure 14.65**

72 Explain how an aerofoil on a racing car can enable it to travel faster around corners.

73 A small hydroelectric power station has an output of 36 kW and an efficiency of 74 per cent. If the water falls a vertical distance of 62 m, determine:

- **a** the mass of the water passing through the turbines every second
- **b** the speed of the water as it reaches the turbines.
- **c** State two assumptions you made during this calculation.

74 Discuss which sports use balls that have trajectories that are most affected by spin, and how this is achieved.

75 Water initially flowing at a speed of $8.0\,\text{cm}\,\text{s}^{-1}$ passes through an arrangement similar to that shown in Figure 14.64. In the narrow section the radius of the tube has been reduced from 3.2 cm to 0.48 cm. This results in a pressure drop observed as a difference in height in the vertical tubes of 58 cm.

- **a** What is this pressure difference expressed in pascals?
- **b** What is the flow speed in the narrow section?
- **c** Calculate the pressure difference if the initial flow rate is decreased to $3.7\,\text{cm}\,\text{s}^{-1}$.

76 Find out how the Bernoulli effect enables a sailing boat to move towards the direction from which the wind is coming.

77 The Venturi effect has many different applications. Research one use of this effect and prepare a short presentation for the rest of your group.

78 An aircraft is flying at a height where the atmospheric pressure is measured to be 7.5×10^4 Pa.

- **a** Use the internet to find out the approximate altitude of the plane.
- **b** A Pitot tube on the aircraft directed towards the front of the plane recorded a pressure of 5.6×10^4 Pa. Calculate a value for the air speed of the aircraft if the density of air at that height is $0.88\,\text{kg}\,\text{m}^{-3}$.

Viscosity

Of course, no liquid is perfectly 'ideal' because there will always be some frictional forces between different layers, and the outer layers and any container.

Viscosity can be considered as a measure of a fluid's resistance to flow.

The (coefficient of) viscosity is given the symbol η (eta) and has the SI unit of Pa s. (This unit is not in common usage; *poise* is more widely used but is not required for this course). Some typical values at 20°C are given in Table 14.2. Viscosities can be very dependent on temperature.

Table 14.2 Some typical viscosities

Fluid	Viscosity, η/Pa s
'heavy' oil	0.7
'light' oil	0.1
water	1×10^{-3}
human blood	4×10^{-3}
gasoline (petrol)	6×10^{-4}
air	2×10^{-5}

Stokes's law

When an object moves through a fluid it will experience a resistive force because of the viscosity of the fluid. This force is known as **viscous drag**, F_D. *Stokes's law* provides a way of calculating the size of this drag force, but only under certain conditions:

- there must be streamline (laminar) flow
- using spherical objects
- with smooth surfaces.

$$F_D = 6\pi\eta r v$$

This is Stokes's law. It is given in the *Physics data booklet*; r is the radius of the sphere, which is moving with a speed, v. Because of its limitations, it is usually only applied to small spherical masses falling under the effects of gravity.

Worked example

21 Calculate the viscous drag force acting on a small metal sphere of radius 2.3 mm falling through air of viscosity 1.7×10^{-5} Pa s at a speed of 38 cm s^{-1}.

$$F_D = 6\pi\eta r v = 6\pi \times (1.7 \times 10^{-5}) \times (2.3 \times 10^{-3}) \times 0.38$$
$$= 2.8 \times 10^{-7}\,\text{N}$$

Terminal speed of a falling sphere

Dropping small spheres through fluids is a widely used method for determining their viscosities and how they may depend on temperature. A method is shown in Figure 14.66 in which an electronic timer is started and stopped as the metal ball passes through the two light gates.

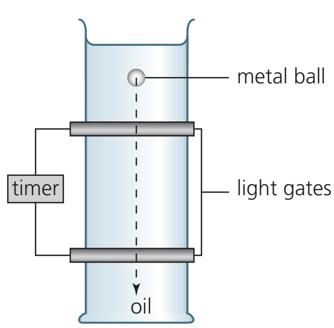


Figure 14.66 Experiment to determine the viscosity of a liquid

The light gates must be placed so that a sphere has reached its terminal speed before passing through the first light gate. Smaller balls will reach their terminal speed in a shorter distance. The tube should be as wide as possible to encourage streamlined flow.

If a sphere (of mass m and radius r) is in equilibrium moving with a constant speed, which we saw in Chapter 2 is called its *terminal speed*, v_t, then the upwards and downwards forces on it are balanced as shown in Figure 14.67.

viscous drag, F_D + buoyancy force, B = weight, mg

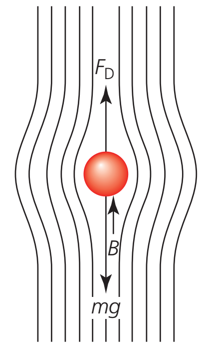
$$6\pi\eta r v_t + \rho_f V_f g = mg$$

but,

$$V_f = \frac{4}{3}\pi r^3$$

so:

$$6\pi\eta r v_t + \rho_f \frac{4}{3}\pi r^3 g = mg$$



■ **Figure 14.67** Streamlines around a sphere falling with terminal speed

If the mass and radius of the sphere are measured and the terminal speed determined as shown in Figure 14.66, then the viscosity of the liquid can be determined if its density is known. Repeating the measurements of terminal speed with spheres of different radius (same material) will enable a graph to be drawn, which will help improve the uncertainties in the determination of viscosity.

If the experiment is repeated with spheres that have double the radius, the viscous drag will be doubled, but the weight and buoyancy force will both be increased by a factor of 2^3 (because they are proportional to volume).

Worked example

22 In an experiment similar to that shown in Figure 14.67, a sphere of radius 8.9 mm and mass 3.1 g reached a terminal speed of 34 cm s^{-1} when falling through an oil of density 842 kg m^{-3}. Determine a value for the viscosity of the liquid.

$$6\pi\eta r v_t + \rho_f \frac{4}{3}\pi r^3 g = mg$$

$$(6 \times \pi \times \eta \times 8.9 \times 10^{-3} \times 0.34) + \left(842 \times \frac{4}{3} \times \pi \times (8.9 \times 10^{-3})^3 \times 9.81\right) = 3.1 \times 10^{-3} \times 9.81$$

$$\eta = 0.11 \text{ Pa s}$$

79 The viscosity of maple syrup is about 3200 times larger than that of water. Determine the density of maple syrup if the terminal velocity of a 5.5 g sphere of radius 6.1 mm falling through the syrup is 11 cm s^{-1}.

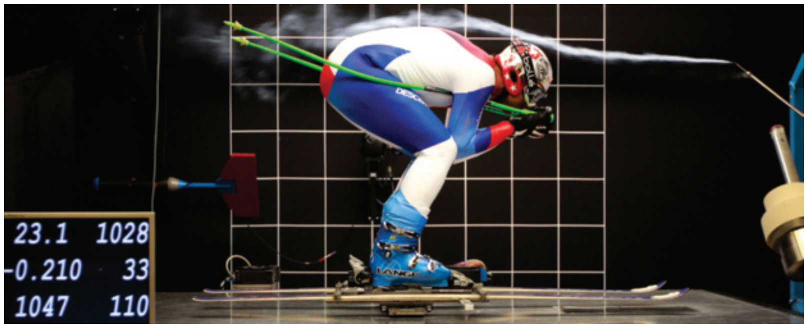
80 **a** Determine the viscous drag force acting on a raindrop of radius 1.0 mm falling at a speed of 4.0 m s^{-1} (viscosity of air = 1.7×10^{-5} Pa s).
b What is the theoretical terminal speed of this falling raindrop? Assume the density of air is 1.2 kg m^{-3}.
c Suggest why the actual terminal speed will be a lot less than the value calculated in (b).

81 In an experiment similar to that shown in Figure 14.66 the terminal speeds of metal spheres were measured for different radii (same metal). Sketch a graph of the possible results.

82 Suggest why the viscous drag represented by Stokes's law is proportional to the radius, *r*, but, more generally, drag forces in fluids are described as being proportional to cross-sectional area (that is, proportional to r^2).

■ Laminar and turbulent flow and the Reynolds number

If streamlines can be observed in a fluid (for example, by using a coloured dye injected into water), turbulence can be investigated by gradually increasing the flow rate until the streamlines become disrupted. The effect of different diameter tubes, or different shaped objects in a fluid flow, can be observed. See Figure 14.68. This is especially important in the *aerodynamic design* of aircraft and other vehicles. Observing computer models of fluid dynamics is recommended.



■ **Figure 14.68** Wind tunnel testing

As the mean speed, v, of a fluid through a pipe of radius r increases, laminar flow becomes less likely and turbulence may begin. The *Reynolds number, R*, is used as a guide to predict the conditions under which turbulent (non-laminar) flow will begin.

$$R = \frac{vr\rho}{\eta}$$

This equation is given in the *Physics data booklet*; R is *dimensionless* (it does not have a unit).

As a generalized guide, if $R < 1000$ we can expect laminar flow.

But it should be understood that fluid flow can be unpredictable and it is easily affected by diverse influences. This number is therefore just a guide, and the onset of turbulence may well depend on other factors not included in the equation – for example, the shape of the container (or moving object) and whether the fluid is being disturbed in any other way.

For non-turbulent flow it is required that:

$$1000 > \frac{vr\rho}{\eta}$$

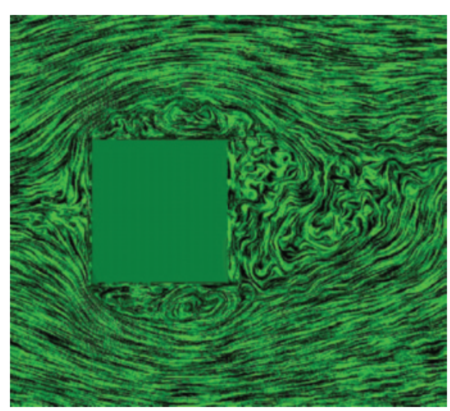
or:

$$v < \frac{1000\eta}{r\rho}$$

This equation demonstrates that a fluid can flow faster without causing turbulence if it has a larger viscosity and smaller density, and if larger dimensions are involved. The same equation can be used for estimating the maximum speed of an object of dimension r through a stationary fluid before turbulence will begin. But remember that this is just a guide. As an extreme example, the Reynolds number for a whale had been estimated to be much bigger than 10^6, which means that it is capable of high speed without causing turbulence.

Figure 14.69 and Figure 14.70 show examples of turbulence. In Figure 14.70 the flow of air is from the left and the turbulence occurs mostly behind the obstacle. Turbulence is often

■ **Figure 14.69** Turbulent flow behind a boat



■ **Figure 14.70** Turbulent flow and eddies around a square object in a wind tunnel characterized by swirling currents, involving some flow in the opposite direction to most of the fluid. These are known as *vortices* and *eddies* (singulars: vortex and eddy). The alternate formation of vortices can result in oscillation forces on an object situated in a fluid flow, which can give rise to resonance effects (see section 14.4).

Worked example

23 Estimate the maximum speed for laminar flow of water through a tube of radius 10 cm (viscosity of water = 1.0×10^{-3} Pa s).

$$v < \frac{1000\eta}{r\rho}$$
$$< \frac{1000(1.0 \times 10^{-3})}{0.1 \times 1.0 \times 10^{3}}$$
$$< 1.0 \times 10^{-2}\,\text{m s}^{-1}$$

Additional Perspectives

Blood flow in the human body

The aorta is the largest artery in the human body. It carries blood away from the heart and downwards to all parts of the body. See Figure 14.71.

A healthy adult heart beats about 70 times every minute, and each beat of the heat pumps a volume of about 70 ml, so that the volume flow rate through the aorta is about $5\,\text{l}\,\text{min}^{-1}$. A volume of 1 ml is the same as 1cm^3, so that $5\,\text{l}\,\text{min}^{-1} = \left(\frac{5000}{60}\right)\text{cm}^3\,\text{s}^{-1} = 83\,\text{cm}^3\,\text{s}^{-1}$. Because:

volume per second = average speed × cross-sectional area

speed of blood in aorta = $\frac{83}{\pi(0.8)^2} = 41\,\text{cm}\,\text{s}^{-1}$ (in an aorta of radius 0.8 cm)

Blood flow in the body is usually streamlined and the conditions in the aorta suggest that turbulence would begin if $R = \frac{vr\rho}{\eta} > 2000$ (approximately). That is, if:

$$v > \frac{2000\eta}{r\rho} = 2000 \times \frac{4 \times 10^{-3}}{0.008 \times 1.05 \times 10^{3}} \approx 1\,\text{m}\,\text{s}^{-1}$$

This value is significantly above typical actual flow rates, so turbulence should not occur.

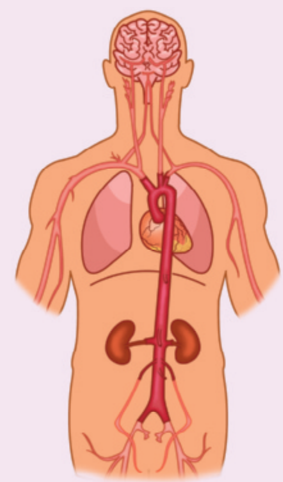


Figure 14.71 The aorta in the human body

Scale models

The flow of a fluid (usually water or air) past a large structure can be investigated in advance of construction by using scale models in wind tunnels. However, if the linear dimensions are reduced in a scale model, the speed at which turbulence is predicted to begin will also change unless the values of viscosity and density of the fluid used in the testing processes are adjusted.

83 Estimate the Reynolds number for a raindrop falling at a speed of $10\,\text{m}\,\text{s}^{-1}$. Is the flow streamlined or turbulent?

84 Estimate the maximum speed of streamlined flow for oil of density $840\,\text{kg}\,\text{m}^{-3}$ and viscosity 0.5 Pa s flowing through a pipeline of radius 25 cm.

85 A design for a new bridge was tested using a scale model in a wind tunnel, with all linear dimensions reduced by a factor of 50. In order for the maximum flow speed observed with the model to be applicable to the full-scale bridge, the density and viscosity of the fluid used in the model needed to be adjusted. If the gas used had a viscosity of half that of air, what density of gas would be required? Assume that the same Reynolds number applies to both the full-scale bridge and the model.

14.4 (B4: Additional Higher) Forced vibrations and resonance *– in the real world, damping occurs in oscillators and has implications that need to be considered*

Many objects are exposed to vibrations from external sources (for example, traffic moving over a bridge could cause it to shake). In this section we will discuss the circumstances under which this effect may be significant, and whether it will be damaging or advantageous in various systems.

Natural frequency of vibration

When something is disturbed and then left to oscillate without further interference, it is said to oscillate at its *natural frequency* of (free) vibration.

In Chapter 4 we discussed the natural frequencies of simple oscillating systems such as pendulums and masses on springs. The energy *dissipation* in such systems may not be extensive, so that they continue to oscillate for some time after they have been disturbed. We will see later that this is described as having a large *Q factor*. Figure 14.72 shows a further example – a hacksaw blade (or a ruler) will vibrate at a natural frequency that depends on its dimensions and the material from which it is made. The frequency can be decreased by adding a load to the blade, for example by taping a mass on the end.

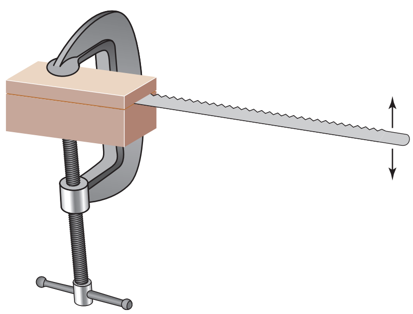


Figure 14.72 Vibrating hacksaw blade

When any object is struck by an external force it will then tend to vibrate naturally, although for most objects the vibrations may be insignificant and/or very short-lasting. Such vibrations can disturb the air around them and send longitudinal waves into the environment, which may be heard as sound.

The two-dimensional standing wave patterns of a horizontal metal plate can be observed by placing small grains (such as fine sand or salt) on a surface that is disturbed into vibration. See Figure 14.73. (Oscillations can be easily encouraged using a mechanical vibrator driven by a signal generator.)

Figure 14.73 Demonstrating the vibrations of a metal plate

An object made of only one material in a simple shape, a tuning fork for example, may produce a single, 'pure', natural frequency, but objects with more complicated structures will have a range of natural frequencies, although one frequency may dominate.

Damping

The motions of all objects have frictional forces of one kind or another acting against them. Frictional forces always act in the opposite direction to the instantaneous motion of an oscillator and result in a reduction of speed and the transfer of kinetic energy (and, consequently, potential energy).

Therefore, as with all other mechanical systems, useful energy is transferred from the oscillator into the surroundings (dissipated) in the form of thermal energy and sound.

Consequently, an oscillator will move at slower and slower speeds, and its successive amplitudes will decrease in size. This effect is called **damping**.

Damping is the dissipation of the energy of an oscillator due to resistive forces.

It is usual for the frequency (and time period) to remain approximately the same during damping as shown in Figure 14.74 because, although the displacements are less, the speeds and accelerations also decrease.

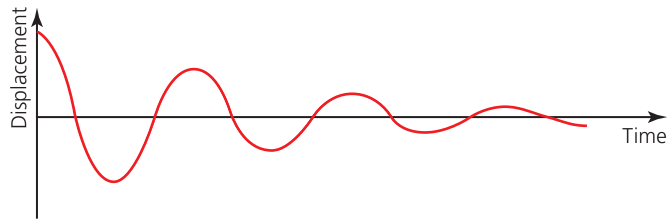


Figure 14.74 Decreasing amplitude (at constant frequency) of a damped oscillation

Damping can be investigated experimentally using simple apparatus like that shown in Figure 14.72 with cards of different areas taped to the blade to increase air resistance (or, for example, using a microphone connected to an oscilloscope to observe the output from a damped tuning fork). It will be found that an approximately exponential relationship exists between amplitude and time.

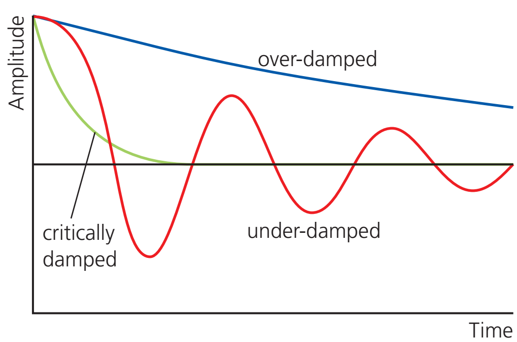


Figure 14.75 Under-damping, over-damping and critical damping

Degrees of damping

The amount (*degree*) of damping in oscillating systems can be very different, as shown in Figure 14.75.

- Some oscillations are **over-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 period) to return to its equilibrium position. The decrease in amplitude with time is often exponential.
- Conversely, occasionally damping can be very light and the oscillator may continue to oscillate, taking some time to dissipate its energy. A pendulum and a mass oscillating on a spring are good examples of **under-damped** systems. If the mass on the spring was placed in a beaker of oil (instead of air) it may then become over-damped.
- Oscillations are often unhelpful or destructive and we may want to stop them as soon as possible. If an oscillation is stopped by resistive forces, such that it settles relatively quickly (compared to its natural period) back into its equilibrium position, without ever passing through it, the process is described as **critically damped**. A car's suspension (see Figure 14.76) is an example of this kind of damping, as are self-closing doors.

Figure 14.76 A shock absorber (damper) on a car is designed to dissipate energy

Utilizations

Taipei 101

One of the world's tallest buildings, Taipei 101, was completed in Taiwan in 2004. Its height, measured to the top of its spire, is 509 m and, at the time of completion, it was the first building in the world to be more than half a kilometre tall.

Taiwan is in a region of the Earth that suffers from the effects of earthquakes and typhoons, so engineers had to be sure that their design could withstand the worst that could happen. A major design feature is a 730 tonne steel pendulum suspended inside the building from the 92nd to 88th floors. This is the major part of a system designed to dissipate energy in the event of strong oscillations produced by typhoons or earthquakes. It is the largest 'damper' in the world.

Without damping, wave energy from the earthquake could be transferred continually during the earthquake, leading to an increasing amplitude of vibration in the building, which could be destructive.

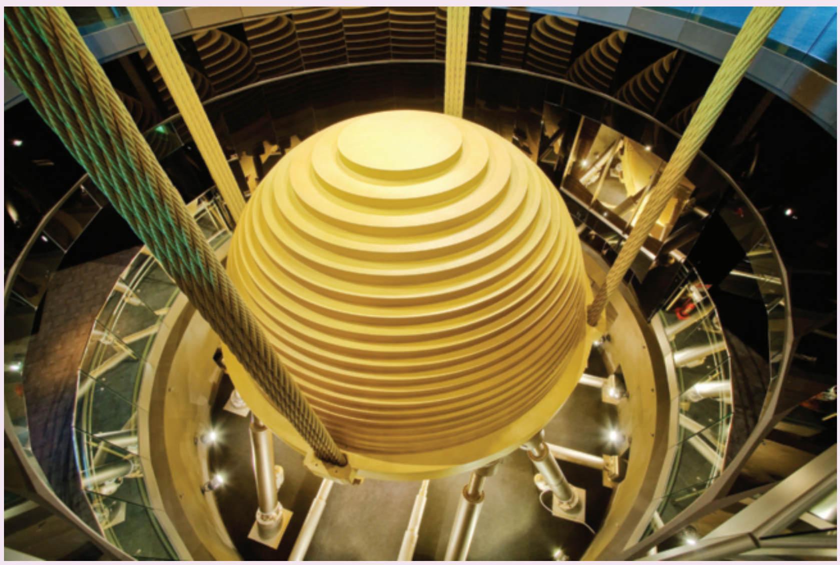


Figure 14.77 The Taipei 101 damper is designed to absorb energy in the event of an earthquake

1. Estimate how much energy could be safely transferred to this pendulum.
2. Very tall buildings usually give engineers many design problems and they are very expensive to build, so why do we build such tall structures? Suggest what kind of limitations there might be on the height of a building.

86 Discuss what factors may affect the natural frequencies of tall buildings.

87 Estimate the natural frequency of your leg when it swings like a pendulum.

88 **a** Suggest how it is possible to dampen the vibrations of a tuning fork.
b Will your suggestions also affect its frequency?

89 A particular guitar string has a first harmonic of 256 Hz.
a What other natural frequencies may occur?
b How can the natural frequency of this string be increased?

90 Explain why the sand in Figure 14.73 demonstrates the standing wave pattern in the plate.

Forced oscillations

We are surrounded by a wide range of oscillations and it is important to consider how these oscillations affect other things. In other words, what happens when an oscillating force is continuously applied to another system? To understand this, it is helpful to first consider a very simple example – what happens when we keep pushing a child on a swing (see Figure 14.78)?

In this case it is fairly obvious – it depends on *when* we push the swing and in *what direction*. If we want bigger swings (increasing amplitudes) we should push once every oscillation in the direction the child is moving. 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.

Figure 14.78 How can we increase the amplitude of a swing?

A **forced oscillation** occurs when an external oscillating force acts on another system tending to make it oscillate at a frequency that may be different from its natural frequency.

Periodic stimulus and the driving frequency

The most important examples of forced oscillation are those in which the frequency of the external force (often called the **driving frequency** or the *forcing frequency*) is the same as the natural frequency. The child on the swing described above is an example of this. When a regular **periodic stimulus** to a system results in an increasing amplitude the effect is called *resonance*.

Resonance

Resonance is the increase in amplitude and energy of an oscillation that occurs when an external oscillating force has the same frequency as the natural frequency of the system.

The oscillations of the driving force must be in phase with the natural oscillations of the system.

Resonance tends to cause oscillations to be amplified. Sometimes the increase in amplitude and the transfer of energy can be considerable.

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/forcing frequency affects the amplitude. 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 14.79 shows a typical arrangement. The resonant frequency of the mass/spring system can be changed by using different springs and/or different masses. The forcing frequency is provided by the electromagnetic vibration generator, which can be driven using different frequencies from the signal generator.

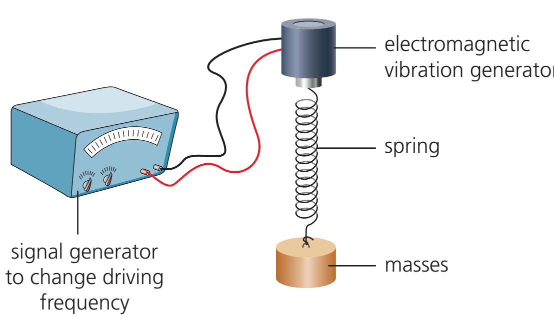


Figure 14.79 Investigating the resonance of a mass on a spring

A typical **frequency–response curve** (see Figure 14.80) rises to a maximum amplitude at the natural frequency, but the sharpness and height of the peak

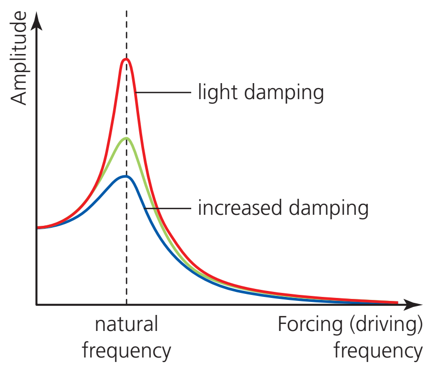


Figure 14.80 Typical frequency–response curves with different degrees of damping

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.

There may be smaller resonance peaks at multiples of the natural frequency (not shown in the diagram).

Worked example

24 The initial amplitude of the damped oscillations of a hacksaw blade was 3.8 cm. After 2.5 s the amplitude had reduced to 2.7 cm.

a Predict the amplitude after 10 s.

b Estimate the time when the amplitude will be 0.5 cm.

a If the change in amplitude is exponential, it will decrease by the same factor in equal time periods. In this example the amplitude after any 2.5 s period can be determined by dividing the amplitude at the start of the period by 3.8/2.7 = 1.41. After 5 s the amplitude will be 1.9 cm, after 7.5 s it will be 1.4 cm, and after 10 s it will be 0.97 cm.

b After 12.5 s the amplitude will be 0.69 cm and after 15 s it will be 0.49 cm, so approximately 15 s are needed for the amplitude to reduce to 0.5 cm (using $A = A_0 e^{-kt}$ produces answers of 0.97 cm and 15 s to two significant figures).

Q factor and damping

We have seen that the amplitude, *A*, of a damped oscillator decreases approximately exponentially with time.

The *Q (quality) factor* of an oscillator is a way of representing the degree of damping involved. A high *Q* factor means that there is little damping. *Q* factor is defined as:

$$Q = 2\pi \frac{\text{energy stored in oscillator}}{\text{energy dissipated per cycle}}$$

One *cycle* means one oscillation. This equation is given in the *Physics data booklet*. Because the energy dissipated per cycle = average power loss × period = power loss/frequency, we can rewrite the equation in an alternative form *for a resonating system oscillating steadily:*

$$Q = 2\pi \times \text{resonant frequency} \times \frac{\text{energy stored in oscillator}}{\text{power loss}}$$

The *Q* factor is a ratio and therefore has no unit.

Remember from Chapter 4 that the energy of an oscillator is proportional to its amplitude squared.

The amplitude of the under-damped oscillator shown in Figure 14.75 reduces to approximately 50 per cent of its previous value in each oscillation (cycle). Because the energy is proportional to the amplitude squared, the energy is reduced to 25 per cent in each time period.

$$Q = 2\pi \frac{1}{0.75} \approx 8$$

The *Q* factor for *critical damping* is usually quoted to be about 0.5, which suggests that most of the energy of the oscillator is dissipated in much less than one time period.

Table 14.3 shows some approximate values of Q factors.

Table 14.3 Some typical Q factors

car suspension system	1
mass on a spring	50
simple pendulum	500
earthquake	1000
tuning fork, guitar string	2000
atomic clock	10^{11}

Worked example

25 The energy of a simple pendulum reduces to half of its initial value in 103 s. If its time period is 2.3 s, estimate its Q factor.

$$Q = 2\pi \frac{\text{energy stored in oscillator}}{\text{energy dissipated per cycle}} = 2\pi \frac{E_p}{E_p/2} \times \text{number of oscillations to halve energy}$$

$$\text{number of oscillations need for energy to halve} = \frac{103}{2.3} = 44.8$$

$$Q = 2\pi \times 2 \times 44.8 = 560$$

This is only an estimate because the energy dissipated per cycle is a mean value over 45 cycles.

91 a The energy of an oscillator reduces from 5.4 J to 4.3 J in successive cycles. What will its amplitude be after one more cycle?
b What is the quality factor of the oscillator?

92 Estimate the Q factor of the hacksaw oscillations described in the Worked example 24 if the period was 0.86 s.

93 a Write down equations for the elastic potential energy stored in a stretched spring and for the period of a mass oscillating on a stretched spring.
b When a mass of 120 g was oscillating with amplitude 8.0 cm on the end of a stretched string the Q factor was found to be 53. if the force constant was 45 N m^{-1}, how much energy was stored in the system?
c Estimate the number of oscillations before the energy in the system is reduced to 60% of the original value. (Assume power loss is constant.)
d How long will this take?

94 What is the Q factor of an oscillator that dissipates 50 per cent of its total energy every cycle?

95 A system is resonating steadily at a frequency of 24 Hz. If its total energy is 4.3 J, determine the power loss if its Q factor is 180.

Examples of resonance

There are many important examples of resonance. Some are useful but many are unwanted and we usually try to reduce their damaging effects. Avoiding resonance in all kinds of structures is a major concern for engineers and it is an interesting combination of physics theory and practical engineering.

Unwanted resonance

- Parts of almost all engines and machinery (and their surroundings) might vibrate destructively when their motors are operating at certain frequencies. For example, a washing machine may vibrate violently 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 close to the frequencies of the earthquake (Figure 14.82).

- Strong winds or currents can 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 14.81.

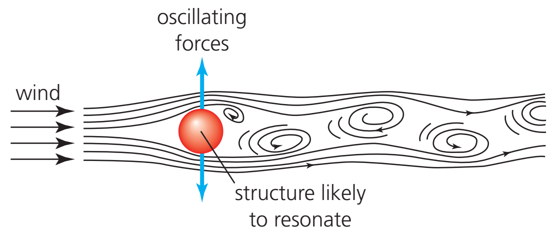


Figure 14.81 A steady wind can cause oscillating forces because of the alternate way in which vortices may be formed

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.

Figure 14.82 Resonance may be one reason why some buildings collapse in an earthquake

Nature of Science

Scientists and engineers must ensure that their structures are safe

Damage that might be caused by resonance (or other reasons) in large buildings and other structures could lead to injuries and death. It is therefore very important that architects and engineers take every possible step to ensure that such damage is extremely unlikely. Regrettably there have been many examples around the world where poor design, materials or workmanship have led to structures collapsing and the loss of life. Mathematical modelling and wind tunnel tests of scale models may be used to test possible circumstances that could affect a structure, but a full-scale structure can never be perfectly represented by models, so further tests are usually carried out on the completed structure.

Utilizations

Bridges

If you have ever crossed a small suspension bridge for walkers (Figure 14.83), 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.

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. More recently, the Millennium Bridge across the River Thames in London had to be closed soon after its opening in June 2000 because of excessive lateral (sideways) oscillations due to resonance (Figure 14.84).

■ **Figure 14.83** Walker on a suspension bridge in Nepal

■ **Figure 14.84** The Millennium Bridge in London was affected by resonance

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 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.

1 Suggest what is meant by 'positive feedback' and give another example.

Useful resonance

- The molecules of certain gases in the atmosphere oscillate at the same frequency as infrared 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 (Chapter 8).
- Radios and TVs around the world are 'tuned' by changing the frequency of an electronic circuit until it matches the driving frequency provided by the transmitted signal.
- 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.

96 Describe a resonance situation in which a high *Q* factor is (a) desirable, (b) undesirable.

97 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.

98 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.

99 Explain in detail how resonance is involved in the greenhouse effect.

Summary of knowledge

14.1 Rigid bodies and rotational dynamics

- From Chapter 6: the angular velocity of an object, $\omega = \frac{\Delta\theta}{\Delta t}$. ($\theta$ is angular displacement, usually measured in radians.) Angular velocity has the unit rad s^{-1}.
- A torque, Γ, is needed to change the motion of an object that is able to rotate.
- The torque provided by a force F, which has a line of action that makes an angle θ with a line joining the point of application of the force to the axis of rotation (length r) can be determined from $\Gamma = Fr\sin\theta$. Torque has the unit N m.
- Torques can be added together to determine the resultant of more than one torque, but their 'direction' (clockwise or anticlockwise) must be taken into consideration.
- An object is in rotational equilibrium if it is rotating with a constant angular velocity (or is at rest). This occurs when there is no resultant torque acting on it.
- A couple is the name given to a pair of parallel, equal magnitude forces that have different lines of action and act in opposite directions, tending to cause rotation.
- Resistance to a change of rotational motion of an object is quantified by its moment of inertia, I, which depends on the distribution of mass around the chosen axis of rotation.
- The moment of inertia of a point mass, m, a distance r from its axis of rotation can be determined from $I = mr^2$. The unit of moment of inertia is kg m^2.
- The moments of inertia of all other masses can (in principle) be found from summing the moments of inertia of all their points: $I = \sum mr^2$. Whenever an equation is needed in an examination for the moment of inertia of a particular shape, it will be provided in the paper.
- A resultant torque will produce an angular acceleration, $\alpha = \frac{\Delta\omega}{\Delta t}$. Angular acceleration has the unit rad s^{-2}. Angular acceleration and linear acceleration, a, are linked by: $\alpha = \frac{a}{r}$.
- The mathematics (including graphs) of linear motion and rotational motion are very similar. By analogy with linear motion we can write down the equations of motion for rotations with constant angular accelerations:
  - $\theta = \left(\frac{\omega_i + \omega_f}{2}\right)t$
  - $\omega_f = \omega_i + \alpha t$
  - $\theta = \omega_i t + \frac{1}{2}\alpha t^2$
  - $\omega_f^2 = \omega_i^2 + 2\alpha\theta$
- The gradient of an angular displacement–time graph is equal to the angular velocity.
- The gradient of an angular velocity–time graph is equal to the angular acceleration.
- The area under an angular velocity–time graph is equal to the change of angular displacement.
- Newton's second law for angular motion: $\Gamma = I\alpha$
- Angular momentum, $L = I\omega$. Unit: kg m^2 s^{-1}
- The total angular momentum of a system is constant provided that no external torques are acting on it.
- The area under a torque–time graph is equal to the change of angular momentum.
- Rotational kinetic energy can be calculated from $E_{K_{rot}} = \frac{1}{2}I\omega^2$. Many objects have both rotational and translational kinetic energy. Rolling down a hill is an important example.
- When an object, such as a wheel, is rolling (without slipping) the point in contact with the surface has a zero velocity at that instant. The translational velocity of a rolling wheel equals the speed of a point on its circumference = ωr.

14.2 Thermodynamics

- Heat engines use the expansion of a gas (when supplied with thermal energy) to do useful mechanical work. Heat engines must work in repeating cycles.
- Changes of state of a gas can be represented on pV diagrams.
- The gas is known as the system; everything else is called the surroundings.
- The equation of state for an ideal gas can be used to predict what happens when a gas changes state: $pV = nRT$.
- The internal energy, U, of an ideal gas is just the translational kinetic energy of its molecules. It can be calculated from $U = \frac{3}{2}nRT$.
- When a fixed mass of gas is compressed or expands, the work done (on or by the gas) can be determined from $W = p\Delta V$, if the pressure is constant. If the pressure varies, the work done can be determined from the area under a pV diagram.
- During a heat engine cycle, each change of state of the gas (system) can usually be approximated to one of four types:
  - Isothermal processes occur at constant temperature, so there is no change to the internal energy of the gas ($\Delta U = 0$); pV = constant (Boyle's law).
  - Isovolumetric processes occur at constant volume, so there is no work done ($W = 0$).
  - Adiabatic processes occur when no thermal energy is transferred to or from the gas ($Q = 0$); $pV^{\frac{5}{3}}$ = constant.
  - Isobaric changes occur at constant pressure.
- The successive stages of a thermodynamic cycle can be represented by characteristic lines on pV diagrams.
- Real changes to a gas may be approximately adiabatic if they are done quickly and are well insulated. Isothermal changes need to be done slowly.
- The principle of conservation of energy can be applied to each change of state: $Q = \Delta U + W$. This is known as the first law of thermodynamics. The meanings of the signs of these three terms need to be clearly understood.
- In a cyclical process, the net work done by a gas is equal to the area enclosed on the pV diagram.
- The efficiency of any process, η = useful work done/energy input.
- The most efficient thermodynamic cycle is known as the Carnot cycle. It involves two adiabatic changes and two isothermal changes, which can be represented on a pV diagram. Thermal energy is absorbed during the isothermal expansion as the gas does work on the surroundings. Thermal energy is released during the isothermal compression.
- The efficiency of a Carnot cycle depends on the temperatures involved: $\eta_{\text{Carnot}} = 1 - (T_{\text{cold}}/T_{\text{hot}})$.
- Left to themselves, ordered things naturally become disordered. Because of the random, uncontrollable nature of uncountable molecular motions and energy transfers, everything that ever happens in the universe increases overall molecular and energy disorder.
- Because of this, useful energy always becomes degraded as it gets dissipated into the surroundings. (It cannot be recovered to do any useful work.)
- Entropy is a numerical measure of the disorder of a system.
- The second law of thermodynamics summarises these ideas – all processes increase the entropy of an isolated system (and the universe as a whole).
- The second law can also be stated as: (i) when extracting energy from a heat reservoir, it is impossible to covert it all into work; (ii) thermal energy cannot spontaneously transfer from a region of lower temperature to a region of higher temperature.
- It is possible to reduce entropy artificially on a local scale, for example by freezing water or growing a plant, but the energy dissipated from the system will always increase the entropy of the surroundings even more.
- The change in entropy of a system, ΔS, when thermal energy ΔQ is added or removed at a constant temperature T(K), can be calculated from $\Delta S = \Delta Q/T$. The units of entropy are J K^{-1}.

■ 14.3 Fluids and fluid dynamics

- A fluid (liquid or gas) exerts a pressure in all directions, including upwards.
- The magnitude of the pressure, p, in a fluid can be determined from the weight of fluid pressing down on an area. This leads to the equation $p = \rho_f gd$, where ρ_f is the density of the fluid and d is the depth of the fluid. This pressure is independent of the shape of any container in which the fluid may be placed.
- If a surface is under two (or more) fluids, the total pressure is found from the sum of the individual pressures. Most commonly, the pressure under a liquid exposed to the atmosphere, $p = p_0 + \rho_f gd$, where p_0 is atmospheric pressure.
- Buoyancy is the ability of a fluid to provide a vertical upwards force on an object placed in, or on, the fluid. This force is sometimes called upthrust.
- The buoyancy force on an object, B, arises because of the difference in pressures between the top and bottom of the object, $B = \rho_f V_f g$. This can be expressed in words – when an object is wholly or partially immersed in a fluid, it experiences an upthrust (buoyancy force) equal to the weight of the fluid displaced.
- Because any liquid is incompressible and its molecular motions are random, we can state the following principle: a pressure exerted anywhere in an enclosed static liquid will be transferred equally to all other parts of the liquid (ignoring any differences caused by changes in depth).
- Hydraulic machinery uses liquids in pipes (often flexible) to transfer pressure. Because $p = F/A$, and the pressure is constant, the magnitude of forces can be increased (or decreased) by choosing different areas.
- A fluid is in hydrostatic equilibrium if it is either at rest, or if any parts of it that are moving have a constant velocity.
- An ideal fluid is incompressible and has a steady, non-viscous flow. It may be visualised as layers of fluid sliding over each other (laminar flow).
- Streamlines are lines that show the paths that (mass-less) objects would follow if they were placed in the flow of a fluid. A tangent to a streamline shows the velocity vector of flow at that point.
- If streamlines get closer together the fluid must have a higher velocity, v. This is also represented in the continuity equation: Av = constant, where A is the cross-sectional area of the enclosed system.
- The Bernoulli equation describes the steady flow of an ideal fluid of density ρ in any system. It includes pressure differences and variations in flow speed, v, and depth, z: $\frac{1}{2}\rho v^2 + \rho gz + p = \text{constant}$.
- For fluids flowing horizontally, the Bernoulli equation reduces to $\frac{1}{2}\rho v^2 + p = \text{constant}$. This shows us that if an (ideal) fluid is flowing horizontally, or an object is moving horizontally through a fluid without turbulence, there must be a decrease in pressure wherever the speed increases. This is known as the Bernoulli effect and it has many applications including aerofoils, spinning balls and Venturi tubes.
- Pitot tubes are used for measuring the speed of fluid flow (for example through a pipe), or the speed of an object through a fluid.
- For a fluid flowing freely out of a hole in an open container, the Bernoulli equation reduces to $\frac{1}{2}\rho v^2 = \rho gz$.
- Viscosity, η, can be considered as a measure of a fluid's resistance to flow. Unit: Pa s
- Stokes's law can be used to calculate a force of viscous drag, F_D. It only applies to smooth, spherical objects experiencing streamlined flow: $F_D = 6\pi\eta rv$.
- Stokes's law is often used to determine viscosity, using data from experiments involving the terminal speed of falling objects.
- The maximum possible speed of laminar fluid flow (or the speed of an object through a stationary fluid) can be predicted using a guide called the Reynolds number, R.
- $R = vr\rho/\eta$. Turbulence will occur if R exceeds a certain value for any particular system. As a generalised guide, if $R < 1000$ we can expect laminar flow.

14.4 Forced vibrations and resonance

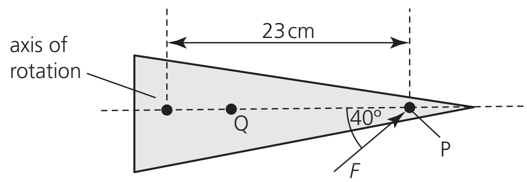
- When they are disturbed, most objects and structures have a natural frequency (or frequencies) at which they will oscillate.
- Frictional forces reduce the speeds and amplitudes of oscillation and dissipate energy. This process is called damping and it can occur to very different degrees in different systems. It may be desirable or unwanted.
- The energy of an oscillator is proportional to its amplitude squared (from Chapter 4).
- Because of damping, successive amplitudes of a system oscillating naturally will get smaller and smaller. The ratio of successive amplitudes may be considered to be constant, so that the decrease in amplitude with time is described as exponential.
- The Q (quality) factor of an oscillator is a way of representing the degree of damping involved: $Q = 2\pi$(energy stored in oscillator/energy dissipated per cycle). This is a ratio and has no unit.
- Critical damping occurs when a system returns relatively quickly to its equilibrium position without passing through it. The Q factor for critical damping is usually quoted to be about 0.5, which suggests that most of the energy of the oscillator is dissipated in much less than one time period.
- Over-damped and under-damped systems can also be identified, and all three types of damping compared on an amplitude–time graph.
- Objects are commonly exposed to forces from external vibrations. If the external force has the same frequency as the natural frequency of the object (and is in phase with it), energy is efficiently transferred and the amplitude increases. This is called resonance.
- The degree of damping (Q factor) determines whether any resonance effects are significant.
- Frequency–response graphs can be drawn to represent resonance in systems with different degrees of damping.
- For a resonating system oscillating steadily, the Q factor can be related to resonant frequency by the equation $Q = 2\pi \times$ resonant frequency $\times$ (energy stored in oscillator/power loss).

■ *Examination questions – a selection*

Paper 3 IB questions and IB style questions

Q1 **a** Explain the meaning of *rotational equilibrium*. (2)

b The diagram shows a vector arrow representing a force acting at point P on a triangular object that is free to rotate about an axis. Determine the value of the force *F* that will produce a torque of 3.4 N m. (2)



c If the object is initially at rest and has a moment of inertia of 1.6 kg m^2, determine the time needed for a constant torque of 2.9 N m to move the object through three complete revolutions. (3)

d Make a copy of the diagram and add to it a vector arrow for any single force acting at Q that will keep the object in rotational equilibrium. (2)

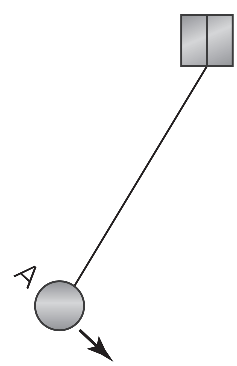
Q2 **a** Explain why a hollow sphere has a greater moment of inertia about an axis through its centre than a solid sphere that has the same material and mass. (2)

b If these two spheres rolled down the same slope from rest, which would have the greater linear speed at the bottom of the slope? Explain your answer. (2)

c Calculate the rotational kinetic energy of a solid sphere rotating at a frequency of 15 Hz if it has a mass of 2.65 kg and a diameter of 16.0 cm. (moment of inertia = (2/5)mr^2) (3)

Q3 **a** Define angular momentum. (1)

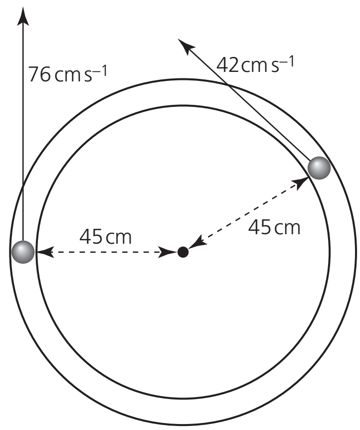
b Consider the oscillations of a simple pendulum that is released from position A, as shown in the diagram.



i Sketch a graph to show how the angular velocity of the pendulum varies with time during the first two complete oscillations. (3)

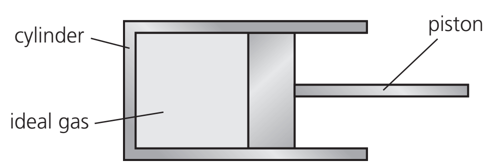
ii The angular velocity of the pendulum varies continuously. Explain why its angular velocity is not conserved. (2)

c Two balls of equal mass are moving at constant speeds in opposite directions on a frictionless horizontal circular track, as shown in the diagram.



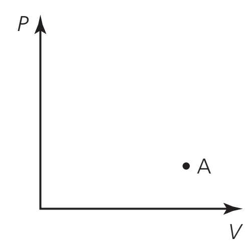
When they meet the collision is a totally (perfectly) inelastic. Determine the magnitude and direction of their *angular* velocities after the collision. (3)

Q4 a An ideal gas is contained in a cylinder by means of a frictionless piston.
At temperature 290 K and pressure 4.8×10^5 Pa, the gas has volume 9.2×10^{-4} m^3.



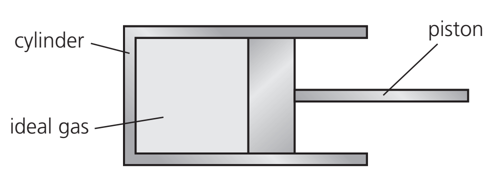
i Calculate the number of moles of gas. (2)
ii The gas is compressed isothermally to a volume of 2.3×10^{-4} m^3. Determine the pressure *P* of the gas. (2)
iii The gas is now heated at constant volume to a temperature of 420 K. Show that the pressure of the gas is now 2.8×10^6 Pa. (1)

b The gas in **a iii** is now expanded adiabatically so that its temperature and pressure return to 290 K and 4.8×10^5 Pa respectively. This state is shown as point A.



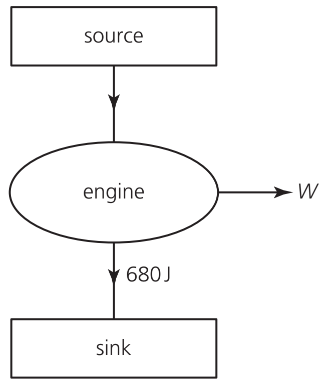
c i Copy the graph and sketch a pressure–volume (*P*–*V*) diagram for the changes in **a ii**, **a iii** and **b**. (3)
ii On your diagram in **c i**, identify with the letter H any change or changes where the gas does external work on its surroundings. (1)
iii Describe how a *P*–*V* diagram may be used to estimate a value for the useful work done in one cycle of operation of an engine. (2)

Q5 a An ideal gas is contained in a cylinder by means of a piston, as shown below.



The piston is pushed quickly into the cylinder.
For the resulting change of state of the gas:
i state, and explain, whether the change is isovolumetric, isobaric or adiabatic (2)
ii use the molecular model of an ideal gas to explain why the temperature of the gas changes. (3)

b A heat engine operates between a high-temperature source and a sink at a lower temperature, as shown below.



The overall efficiency of the engine is 15%. The engine transfers 680 J of energy to the sink.

i Determine the work *W* done by the engine. (2)

ii There is a gain in entropy as a result of the engine doing work *W*. Identify two further entropy changes and, by reference to the second law of thermodynamics, state how the three changes are related. (4)

© IB Organization

Higher Level only

Q6 In March 2012 the movie director James Cameron descended in a vehicle to a depth of almost 11 km below the surface of the Pacific Ocean.

a i Estimate the pressure exerted on the vehicle by the sea water at this depth (density = 1025 kg m^{-3}). (2)

ii Explain why it is unnecessary to include the atmospheric pressure in the previous calculation. (1)

iii What force is acting inwards on each square centimetre of the vehicle's surface? (1)

b i Explain why the vehicle experiences an *upthrust* from the surrounding water. (2)

ii Calculate the upthrust on a vehicle of volume 15 m^3 at this depth. (2)

c Suggest how a submarine is able to change its depth under the surface of the ocean. (1)

Q7 The equation $F_D = 6\pi\eta rv$ (Stokes's law) can be used to determine the viscous drag, F_D, on an object moving through a fluid under certain circumstances.

a i Explain the meaning of the term *viscosity*. (1)

ii Under what circumstances can this equation be applied to the motion of a smooth, spherical object? (1)

b A metal sphere of radius 1.0 mm and density 3.7×10^3 kg m^{-3} falls vertically through oil at a constant speed.

i Determine the buoyancy force on the sphere if the density of the oil is 910 kg m^{-3}. (2)

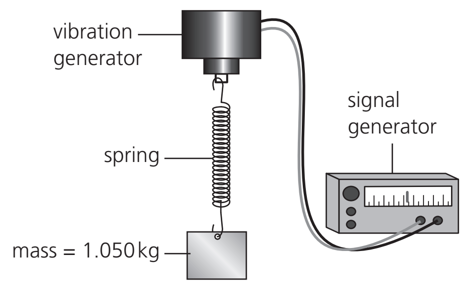
ii Determine the speed of the sphere if the viscosity of the oil is 0.048 Pa s. (3)

c Estimate the minimum speed of the sphere through the oil that would result in turbulence. Assume that turbulence occurs if the Reynolds number is greater than 1000. (2)

Q8 A mass of 1 kg is hanging on a spring that has a force constant of $20\,N\,m^{-1}$.

a Show that the natural frequency of vibration of this system is approximately 0.7 Hz. (1)

b The spring and mass are then vibrated, as shown in the diagram, by a vibration generator. The frequency supplied by the signal generator is varied from 0 Hz to 2 Hz. Sketch a graph to show how the amplitude of the oscillations of the mass changes over this range of frequencies. (3)



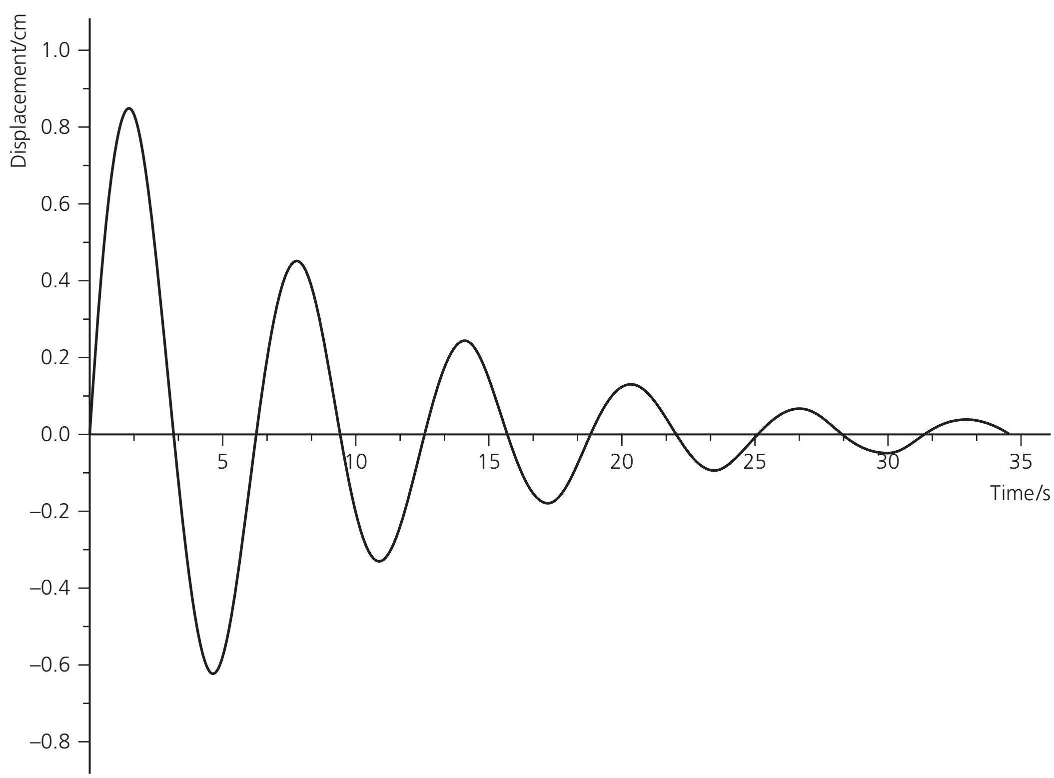
c The experiment is then repeated with the mass immersed in some water in a beaker. On the same axes as your previous graph, draw another graph to show the variation of amplitude of the mass in water when the signal generator is operated at the same power as in **c**. (2)

d Masses vibrating on springs can be used as models for molecular oscillations. Outline how molecular resonance can be used to explain how 'greenhouse gases' can absorb energy radiated from the Earth's surface. (3)

Q9 **a** State what is meant by *damped* oscillations. (2)

b **i** The following diagram shows how the displacement of an oscillating system varies with time. State the type of damping shown. (1)

ii Take measurements from the diagram to determine whether the amplitude of the oscillation decreases exponentially with time. (3)



c Estimate the Q factor of the system. (2)

Option C

15 Imaging

ESSENTIAL IDEAS

- The progress of a wave can be modelled using the ray or the wavefront. The change in wave speed when moving between media changes the shape of the wave.
- Optical microscopes and telescopes utilize similar physical properties of lenses and mirrors. Analysis of the universe is performed both optically and by using radio telescopes to investigate different regions of the electromagnetic spectrum.
- Total internal reflection allows light or infrared radiation to travel along a transparent fibre. However, the performance of a fibre can be degraded by dispersion and attenuation effects.
- The body can be imaged using radiation generated from both outside and inside. Imaging has enabled medical practitioners to improve diagnosis with fewer invasive procedures.

15.1 (C1: Core) Introduction to imaging *– the progress of a wave can be modelled via the ray or the wavefront; the change in wave speed when moving between media changes the shape of the wave*

How we see images

We see an object when light from it enters our eyes. Some objects emit light, but we are able to see most things because the light waves striking them are scattered in all directions and some of the waves spreading away from a particular point on the object are brought back together to a point in our eyes. Figure 15.1 shows this using *rays* to represent the directions in which the waves are travelling. The representation of an object that our eyes and brain 'see' is called an **image**. The term **object** is generally used to describe the thing that we are looking at.

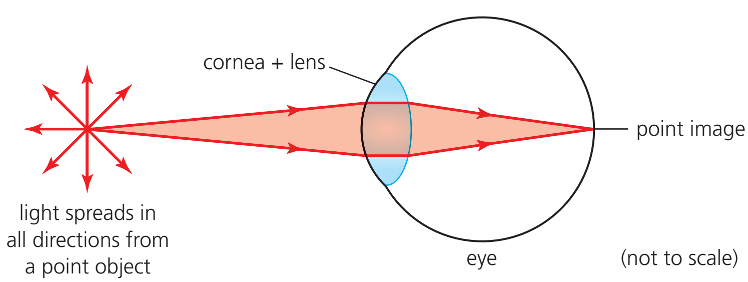


Figure 15.1 The eye focusing light to form an image

The eye uses *refraction* to bring light rays diverging from a point on an object back to a point on the image. This process is called **focusing** the light to form an image.

Additional Perspectives

Understanding the human eye

Figure 15.2 shows the basic structure of the human eye. Light rays are refracted as they pass into the eye through the *cornea*. Further refraction then takes place at surfaces of the *lens*. As a result the rays are focused on the back of the eyeball (*retina*) where the image is formed.

The *iris* controls the amount of light entering the eye. The aperture (opening) through which the light passes is called the *pupil*. In bright light the iris decreases the size of the pupil to protect the eye, while at night the pupil *dilates* (gets larger) so that more light can

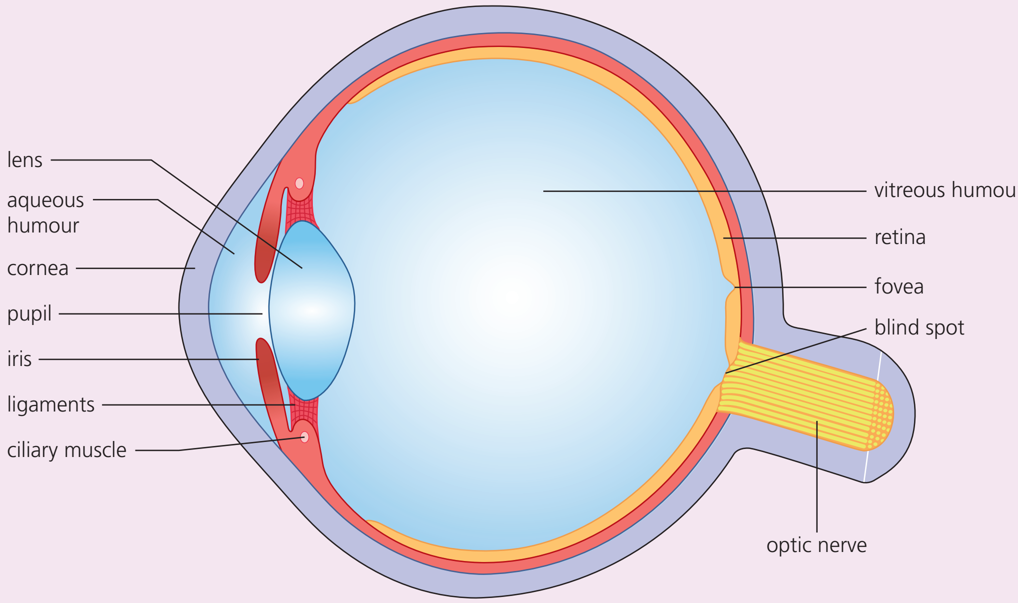


Figure 15.2 Physical features of the human eye

be received by the retina in order to see clearly. The *aqueous humour* is a watery liquid between the cornea and the lens; the *vitreous humour* is a clear gel between the lens and the retina.

The *ciliary muscles* can change the shape of the lens; this is how the eye is able to focus on objects that are at different distances away.

1 If images are not formed on the surface of the retina, the eye will not be able see clearly. Suggest possible reasons why this may happen.

2 Suggest the purpose of the 'optic nerve'.

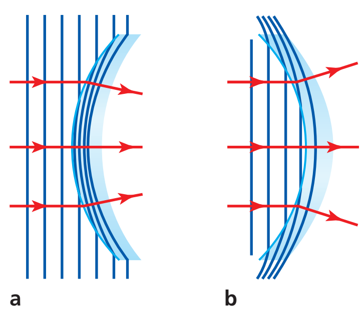


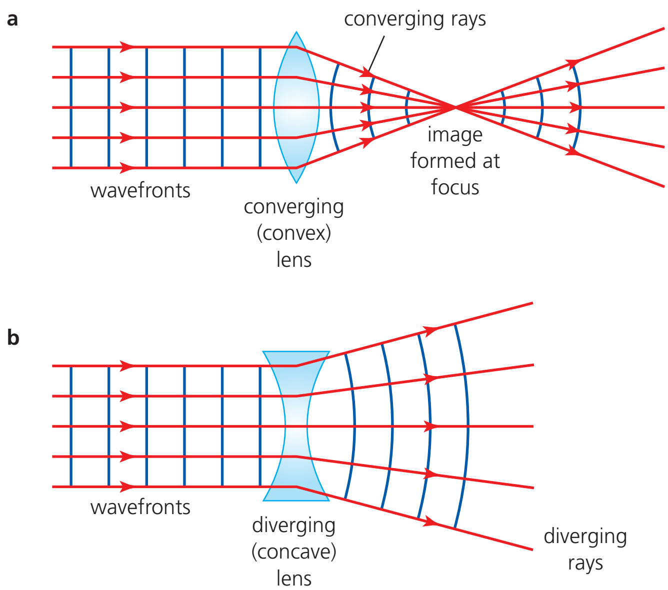
Figure 15.3 How curved interfaces between transparent media affect wavefronts and rays

We know from Chapter 4 that when wavefronts enter a different medium and their speed changes, they can refract and change direction. If the *interface* (boundary) between the two media has a curved surface, then the refracted wavefronts will change shape. Figures 15.3a and 15.3b show plane wavefronts (parallel rays) crossing an interface into a medium where they travel slower. Rays showing the direction of travel of the wavefronts are also included (rays are always drawn perpendicular to wavefronts). In 15.3a the incident waves arrive at a *convex* surface and the transmitted wavefronts and rays *converge*. In 15.3b the waves are incident on a *concave* surface and the transmitted wavefronts and rays *diverge*.

Converging and diverging lenses

The eye contains a *lens* that helps to focus the light. Manufactured lenses made of *transparent* materials (such as glass or plastic) use the effect shown in Figure 15.3 to focus light and form images. This usually involves light travelling from an object through air and then through a transparent lens that has two smooth, curved surfaces. Refraction then occurs at both surfaces as shown in Figure 15.4, which shows the effects of the two basic types of lens on plane wavefronts. The wavefronts inside the lenses have not been included in these diagrams. Light rays will refract and change direction at both surfaces of the lens, unless they are incident along a normal. However, in the rest of this chapter we will usually simplify the diagrams in order to show the rays changing direction only once – in the centre of the lens.

■ **Figure 15.4** Two basic types of lens and how they affect light waves (and rays)



In Figure 15.4a the wavefronts converge to a focus–for this reason this type of lens is often called a **converging lens**. Because of the shape of its surface, this type of lens is also called a convex lens. Despite their name, converging lenses do not *always* converge light (magnifying glasses are the exception). Figure 15.4b shows the action of a **diverging lens** (concave surface). Lenses are made in a wide variety of shapes and sizes, but all lenses can be described as either converging/convex or diverging/concave.

Lenses have been in use for thousands of years in many societies around the world. The oldest were crafted from naturally occurring translucent rock (see Figure 15.5). They may have been used for magnification or for starting fires.

■ **Figure 15.5** The oldest known lens (found at the Assyrian palace at Nimrud); it is now in the British Museum in London

Thin lenses

Although real lenses will not behave exactly as the idealized descriptions and equations presented in this chapter, lens theory can be applied confidently to *thin lenses* (which have surfaces with small curvatures) and for light incident approximately perpendicularly (*normally*) close to the middle of such lenses.

Terminology

Figure 15.6 illustrates the basic terms used to describe lenses.

Figure 15.6 Defining the basic terms used to describe lenses

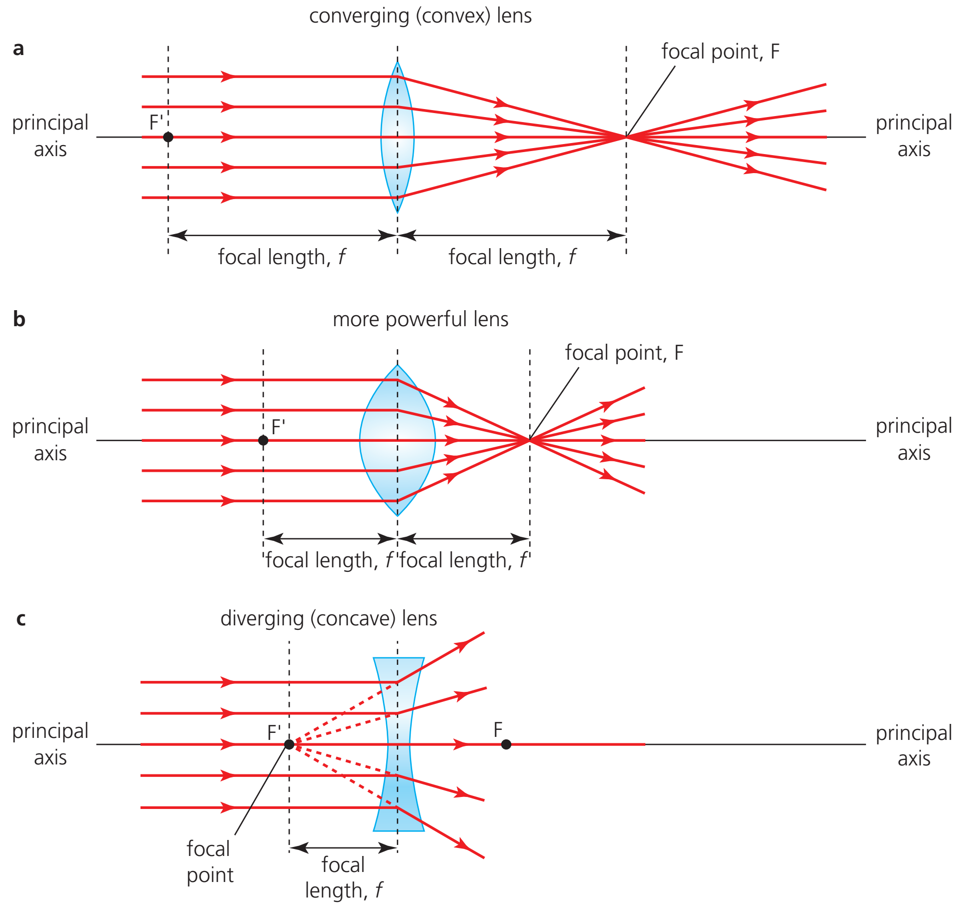
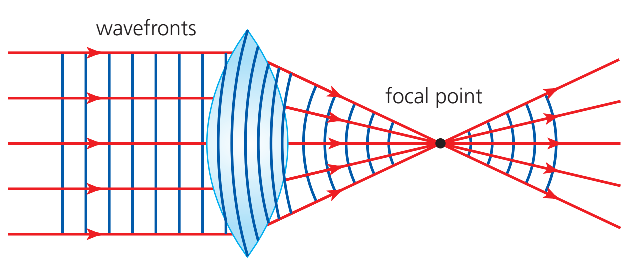


Figure 15.6 shows ray diagrams, and for the rest of this chapter we will continue to use rays because they are usually the easiest way of representing the behaviour of optical systems. However, as an example, Figure 15.7 shows how the behaviour of the converging lenses shown in Figure 15.6b could be represented using wavefronts.

Figure 15.7 Wavefronts being focused by a converging lens



The **principal axis** of a lens is defined as the (imaginary) straight line passing through the centre of the lens, which is perpendicular to the surfaces.

Light rays may be focused in different places depending on how close the object is to the lens, but a lens is defined in terms of where it focuses parallel rays of light that are incident on it.

The **focal point** of a converging lens is defined as the point through which all rays parallel to the principal axis converge, after passing through the lens. For a diverging lens the focal point is the point from which the rays appear to diverge after passing through the lens.

The focal point is sometimes called the *principal focus*. A lens has two focal points, the same distance from the centre of the lens on either side. These are shown as F and F' in Figure 15.6.

The **focal length**, f, of a lens is defined as the distance along the principal axis between the centre of the lens and the focal point.

Focal length is typically measured in centimetres, although the SI unit is metres. The focal length of a lens is the essential piece of information about a lens that tells us how it affects light passing through it. The longer the focal length of a lens, the less effect it has on light. The shorter the focal length of the lens, the bigger the refraction of the light, and the lens is described as being more *powerful*.

For reasons that will be explained later, the focal lengths of diverging lenses are given *negative* values.

To determine the focal length of a lens experimentally it is necessary to use parallel rays of light. These are conveniently obtained from any distant object – spherical wavefronts from a point source become effectively parallel if they are a long distance from their origin.

The focal length of a lens depends on the curvature of the surfaces and the refractive index of the material(s) from which the lens is made. Simple lenses have surfaces that are spherical – the same shape as part of a sphere. A lens with a smaller radius of curvature, or a higher refractive index, will have a shorter focal length and be more powerful (see Figure 15.6b). Eyes are able to focus objects at different distances away by slightly changing their shape and, therefore, their focal lengths (a process called *accommodation*).

People who work with lenses, such as optometrists and opticians, usually classify different lenses according to their (optical) power, a term that is not connected in any way to the more general meaning of power as the rate of transfer of energy. **Optical power** is defined by:

$$\text{power} = \frac{1}{\text{focal length}}$$

$$P = \frac{1}{f}$$

This equation is given in the *Physics data booklet*.

The unit for (optical) power is the **dioptre**, D, which is defined as the power of a lens with a focal length of 1 m. That is:

$$P\,(\text{D}) = \frac{1}{f\,(\text{m})}$$

When two lenses are placed close together, their combined power is equal to the sum of their individual powers.

Worked example

1 What are the powers of lenses that have focal lengths of:
a +2.1 m
b +15 cm
c −50 cm?

a $P = \frac{1}{f} = \frac{1}{2.1} = +0.48\,\text{D}$

b $P = \frac{1}{f} = \frac{1}{0.15} = +6.7\,\text{D}$

c $P = \frac{1}{f} = \frac{1}{-0.5} = -2.0\,\text{D}$

1 **a** What is the focal length of a convex lens with a power of +2.5 D?
b Make a sketch of a lens (of power +2.5 D) and then next to it draw a lens of the same diameter that has a much shorter focal length.
c What assumption did you make?

2 **a** Calculate the power of a lens with a diameter of 4.0 cm and a focal length of 80 mm.
b How is it possible that another lens of exactly the same shape could have a focal length of 85 mm?

3 A lens of what focal length can be combined with a lens of power +5 D to make a combined power of +25 D?

4 A pair of reading glasses (spectacles) have a power of +1.5 D.
a What kind of lenses do they contain?
b What is the focal length of the lenses?
c If the focal length of the focusing system in the eye is +18 mm, what is the combined power of the eye and the reading glasses?

5 Make large copy of Figure 15.6c and then add wavefronts passing through the system.

Forming images with converging lenses

The properties of an image formed by a converging lens can be investigated using an illuminated object and moving a screen (and/or the object) until a well-focused image is observed. Variations in the image can be observed as the lens is moved, or if the lens is exchanged for another with a different focal length.

The properties of an image

An image can be fully described by listing these properties:

- its position
- whether it is upright or inverted (the same way up as the object or upside down)
- its size (and whether it is magnified or diminished)
- whether it is *real* or *virtual*.

Real and virtual images

Real images are formed where rays of light actually converge. Virtual images are formed when diverging rays enter the eye and the image is formed where the rays appear to have come from. (For example, the images seen when looking at ourselves in a plane mirror or using a magnifying glass are virtual.)

Nature of Science

Deductive logic

By definition, a virtual image cannot be observed directly. Knowledge about virtual images must come from logical reasoning and assessment of other known facts (by *deduction*). Because it is evidently true that (real) images are formed where rays that originated at a point on an object are brought back together at another point, it is logical to conclude that when we can see a virtual image, the image is formed in a similar way (by rays diverging from a virtual point).

Deductive reasoning (logic) produces specific conclusions from generalized true statements. For example, because we know that all forces occur in pairs (from Newton's third law), we can *deduce* that a gun must recoil when it is fired.

Linear and angular magnification

The *magnification* of an image tells us how much bigger or smaller the image is compared to the object, but this can be expressed in two different ways.

Linear magnification, *m*

The *linear magnification*, *m*, of an image is defined as the ratio of the height of the image, h_i, to the height of the object, h_o.

$$m = \frac{h_i}{h_o}$$

This equation is given in the *Physics data booklet*.

Because *m* is a ratio it has no unit. If *m* is larger than one, the image is magnified; if *m* is smaller than one, the image is diminished (smaller).

Angular magnification, M

Sometimes the dimensions of an object and/or an image are not easily determined, or sometimes quoting a value for a linear magnification may be unhelpful or misleading. For example, an image of the Moon that had a diameter of 1 m would be impressive, but its linear magnification would be $m = h_i/h_o = 1/(3.5 \times 10^6) = 2.9 \times 10^{-7}$. In such cases the concept of **angular magnification** becomes useful. See Figure 15.8.

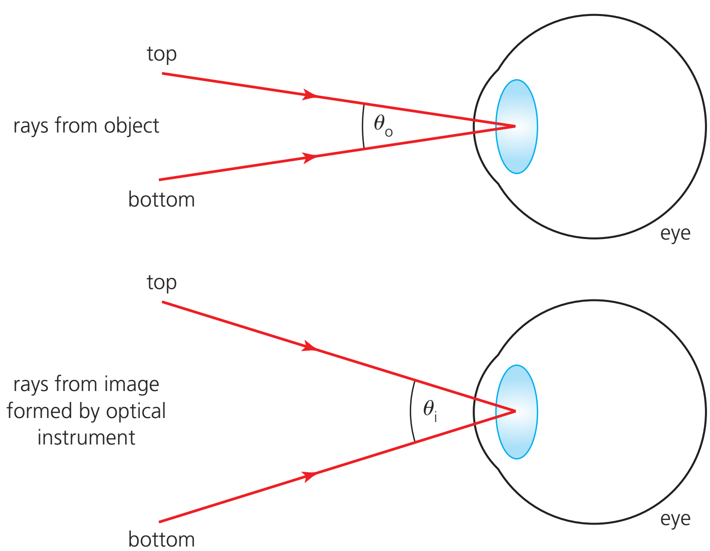


Figure 15.8 The concept of angular magnification

Angular magnification, M, is defined as the angle subtended at the eye by the image, θ_i, divided by the angle subtended at the eye by the object, θ_o. Because it is a ratio, it has no unit.

$$M = \frac{\theta_i}{\theta_o}$$

This equation is given in the *Physics data booklet*.

Returning to the example of an image of the Moon, if a 1 m diameter image of the Moon was viewed from a distance of 2 m, it would subtend an angle of $\frac{1}{2}$ rad (or 29°) at the eye. The Moon is an average distance of 3.8×10^8 m from Earth, so it subtends an angle of $(3.5 \times 10^6)/(3.8 \times 10^8) = 9.2 \times 10^{-3}$ rad at our eyes. The angular magnification is found from $M = 0.50/9.2 \times 10^{-3} = 54$.

6 When a magnifying glass was used to look at a small insect it appeared to have a length 3.7 mm. If the linear magnification of the lens was 4.6, what was the real length of the insect?

7 A picture of width 4.0 cm and height 2.5 cm is projected on to a screen so that it is 83 cm wide.
a What is the linear magnification?
b How tall is the image?
c By what factor has the area of the image increased?

8 The angular magnification of a telescope was 12 when it was used to look at a tree 18 m tall. If the tree was 410 m away, what angle was subtended by the image of the tree at the eye of the observer?

Predicting the properties of real images formed by converging lenses

The position and properties of an image can be predicted theoretically by using one of two methods:

- scale drawing (ray diagrams)
- the (thin lens) equation, which links the object and image positions to the focal length of the lens.

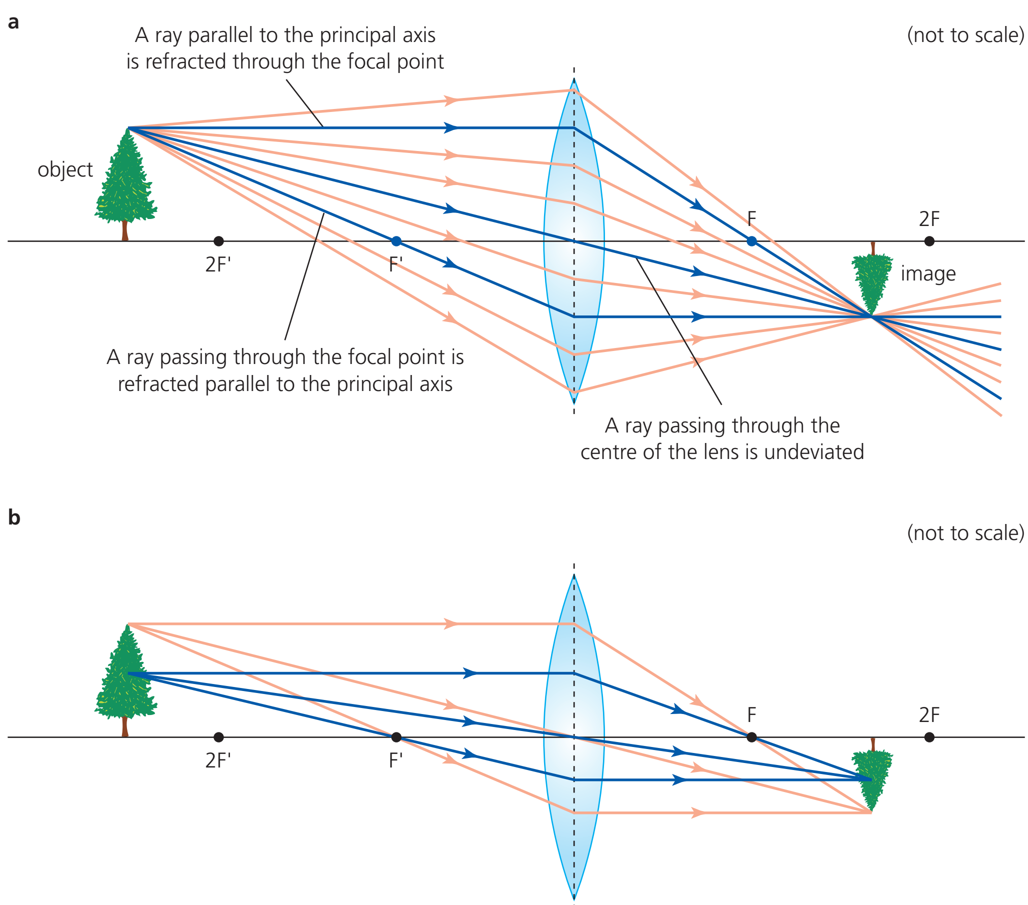
Using ray diagrams

Figure 15.9a shows rays that are coming from the top of an *extended* object (that is, it is not a *point* object) being focused to form an image. All rays incident on the lens are focused to the same point. If part of the lens was covered, an image would still be formed at the same point by the remaining rays.

The predictable paths of three rays coming from the top of the object are highlighted. These same three rays can be used to locate the image in any situation.

- A ray parallel to the principal axis passes through the focal point.
- A ray striking the centre of the lens is undeviated.
- A ray passing through the focal point emerges from the lens parallel to the principal axis.

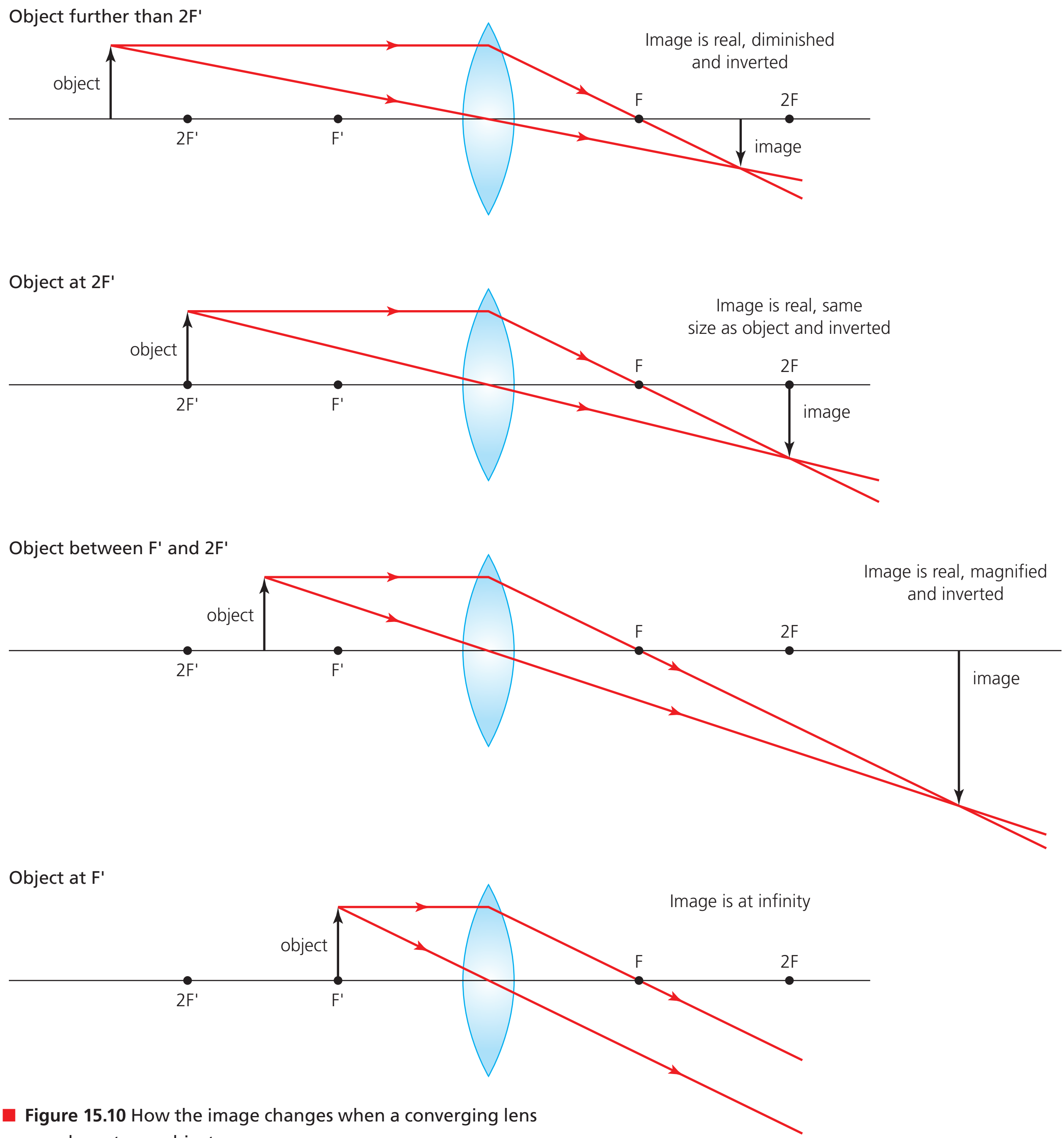
Figure 15.9 Predicting the paths of rays between an object and its image using three standard rays



Note that the vertical scale of the diagram is misleading – a light ray striking the centre of a small thin lens from an object some distance away will be incident almost normally, which is not apparent in this diagram. Of course, all the light from an object does not come from one point at the top. Figure 15.9b also shows the paths of three rays going from the middle of the object to the middle of the image.

In the example shown in Figure 15.9, we can see from the ray diagram that the image is between the positions F and 2F (the point 2F is a distance 2*f* from the centre of the lens) and it is diminished, inverted and real. If the lens shown was replaced by a less powerful lens, the image would be further away, bigger and dimmer (but it would remain inverted and real).

If a lens and object are brought closer together, the image stays real and inverted but becomes larger and further away from the lens (as well as dimmer). But if the object is placed at the focal point, the rays will emerge parallel and not form a useful real image (it is at infinity). Figure 15.10 represents these possibilities in a series of diagrams for easy comparison.



■ **Figure 15.10** How the image changes when a converging lens moves closer to an object

If an object is placed closer to the lens than the focal point, the emerging rays diverge and cannot form a real image. Used in this way, a lens is acting as simple magnifying glass, and a virtual, magnified image can be seen by an eye looking through the lens, as shown in Figure 15.12, which will be discussed later in this chapter.

9 **a** Draw a ray diagram to determine the position and size of the image formed when an object 10 mm tall is placed 8.0 cm from a convex lens of focal length 5.0 cm.
b What is the linear magnification of the image?

10 **a** Draw a ray diagram to determine the position and size of the image formed when an object 20 cm tall is placed 1.20 m from a convex lens of power 2.0 D.
b What is the linear magnification of the image?

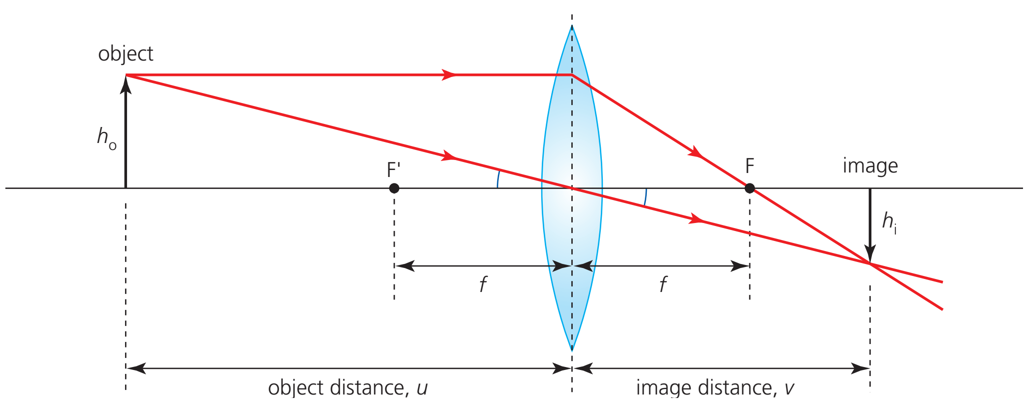
11 Construct a ray diagram to determine where an object must be placed in order to project an image of linear magnification 10 on to a screen that is 2.0 m from the lens.

12 An image of an object 2.0 cm in height is projected on to a screen that is 80 cm away from the object. Construct a ray diagram to determine the focal length of the lens if the linear magnification is 4.0.

13 **a** Describe the properties of images that are formed by cameras.
b Draw a sketch to show a camera forming an image of a distant object.
c How can a camera focus objects that are different distances away?

Using the thin lens equation

The thin lens equation provides a mathematical alternative to scale drawings for determining the position and properties of an image. In this equation the symbol u is used for the distance between the object and the centre of the lens (called the *object distance*) and the symbol v is used for the distance between the image and the centre of the lens (the *image distance*), as shown in Figure 15.11.



■ **Figure 15.11** Object and image distances

The thin lens equation is given in the *Physics data booklet*:

$$\frac{1}{f} = \frac{1}{v} + \frac{1}{u}$$

It is possible to put in values for f and u (when $u < f$) that would lead to a *negative* value for the image distance, v, so we need to understand what that means. A negative image distance means that the image is virtual (we will discuss virtual images again in the next section). More generally, we need to make sure that when inputting data into the thin lens equation we use the correct signs, as summarized in the *'real is positive' convention*:

Real is positive convention

- Converging lenses have *positive* focal lengths.
- Distances to real objects and images are *positive*.

- Upright images have *positive* linear magnifications.
- Diverging lenses have *negative* focal lengths.
- Distances to virtual images are *negative*.
- Inverted images have *negative* linear magnifications.

Looking at the two similar triangles with marked angles in Figure 15.11, it should be clear that:

$$\frac{h_o}{u} = \frac{h_i}{v} \text{ or } \frac{h_i}{h_o} = \frac{v}{u}$$

Therefore, the magnitude of the linear magnification, m, $(= \frac{h_i}{h_o})$ can also be calculated from $\frac{v}{u}$, but a negative sign is added because of the 'real is positive convention':

$$m = -\frac{v}{u}$$

This equation is listed in the *Physics data booklet*.

ToK Link

Conventions

Could sign convention, using the symbols of positive and negative, emotionally influence scientists?

The 'real is positive' convention is used in this course, but there is another widely used alternative (which is not included). There are other situations in physics were we need to decide on a convention (for example, current flowing from positive to negative). And the choice of positive charge for protons and negative for electrons could easily have been the other way around. Provided that everyone understands the convention that is being used, it is not of great significance which system is used, although through cultural influences we may subjectively be inclined to wrongly believe that 'positive' is more important than 'negative'.

Worked example

2 **a** Use the thin lens formula to calculate the position of the image formed by a converging lens of focal length 15 cm when the object is placed 20 cm from the lens.

b What is the linear magnification?

c Is the image upright or inverted?

a $\frac{1}{f} = \frac{1}{v} + \frac{1}{u}$

$\frac{1}{15} = \frac{1}{20} + \frac{1}{v}$

$v = 60\,\text{cm}$

b $m = -\frac{v}{u} = -\frac{60}{20} = -3.0$

c The negative sign confirms that the image is inverted.

Sometimes it is convenient to be able to calculate magnification from simply knowing how far an object is from a lens of known focal length; $m = -(v/u)$ can be combined with the lens equation to show that:

$$m = \frac{f}{u - f}$$

This equation is *not* given in the *Physics data booklet*.

Additional Perspectives

Deriving the thin lens equation

Consider Figure 15.11 again. The ray passing through the focal point on the right-hand side of the lens forms the hypotenuse of two similar right-angled triangles. Comparing these two triangles, we can write:

$$\frac{h_o}{f} = \frac{h_i}{v - f}$$

$$\frac{h_i}{h_o} = \frac{v - f}{f}$$

But we have already that $\frac{h_i}{h_o} = \frac{v}{u}$.

Comparing the two equations, it is clear that:

$$\frac{v}{u} = \frac{v - f}{f}$$

$vf = uv - uf$, or $vf + uf = uv$

Dividing by uvf, we get:

$$\frac{1}{f} = \frac{1}{v} + \frac{1}{u}$$

The important simplifying assumptions made in this derivation are that:

- the ray parallel to the principal axis changes direction in the middle of the lens
- the ray passing through the middle of the lens does not deviate because it is incident normally.

These assumptions are only valid for rays striking a thin lens close to the principal axis.

1 Draw a ray diagram showing the formation of a real image by the refraction of rays at both surfaces of a converging lens.

Use the thin lens formula to answer the following questions about forming real images with convex lenses.

14 In an experiment investigating the properties of a converging lens, image distances were measured for a range of different object distances.
   a Sketch the shape of a graph that would directly represent the raw data.
   b How would you process the data and draw a graph that would enable an accurate determination of the focal length?

15 **a** Determine the position of the image when an object is placed 45 cm from a converging lens of focal length 15 cm.
   b Calculate the linear magnification.

16 **a** Where must an object be placed to project an image on to a screen 2.0 m away from a lens of focal length 20 cm?
   b What is the linear magnification?

17 An object is placed 10 cm away from a converging lens and forms an image with a linear magnification of −3.5. What is the focal length of the lens?

18 What power lens is needed to produce an image on a screen 12 cm away, so that the length of the image is 10 per cent of the length of the object?

19 **a** Derive the equation $m = f/(u - f)$.
   b What focal length of converging lens will produce a magnification of 2 when an object is placed 6.0 cm away from the lens?

The range of normal human vision

The adult human eyeball is between 2 cm and 3 cm in diameter and the focal length of its lens system must be a similar length so that parallel light from distant objects is focused on the back of the eye (the retina).

Muscles in the eye alter the shape of the lens in order to change its focal length (power) so that objects at different distances can be focused on the retina. These muscles are more relaxed when viewing distant objects and most strained when viewing close objects. However, the normal human eye is not powerful enough to focus light from an object that is closer than about 25 cm.

A ray diagram, or the use of the thin lens formula, will confirm that the images formed on the retina are always real, inverted and diminished.

The nearest point to the human eye at which an object can be clearly focused (without straining) is called its **near point**.

The distance from the eye to the near point for a person with normal eyesight (without any aid) is usually assumed to be 25 cm. This distance is often given the symbol *D*.

The furthest point from the human eye that an object can be clearly focused (without straining) is called its **far point**.

A normal eye is capable of focusing objects that are a long way away (although they cannot be seen in detail). The far point is assumed to be at infinity for normal vision.

Simple magnifying glass

In order to see an object in more detail we can move it closer to our eyes, but it will not normally be in focus if the distance to the eye is less than 25 cm. The use of a single converging lens can help to produce a magnified image. Figure 15.12 shows the use of a converging lens as a simple magnifying glass. It produces both an angular magnification and a linear magnification.

Image at the near point

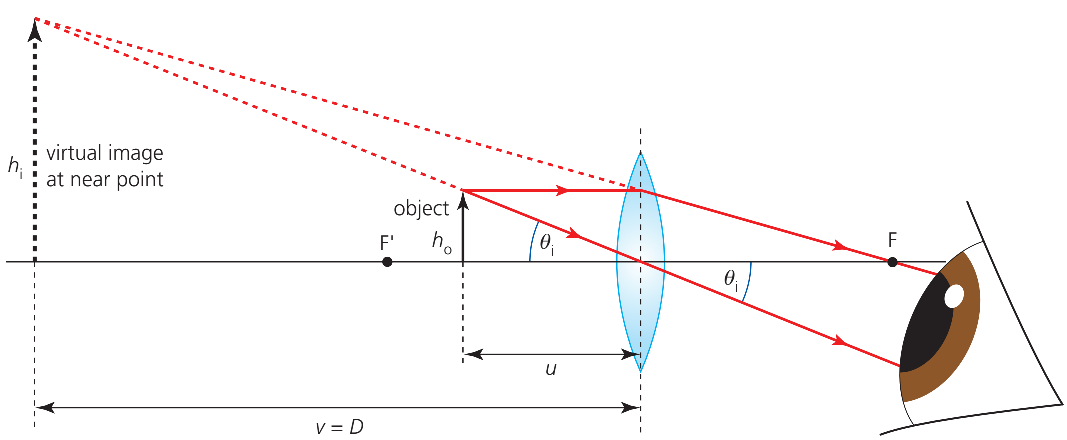


Figure 15.12 A simple magnifying glass forming an image at the near point of the eye (not to scale)

The object must be placed closer to the lens than the focal point, so that the rays diverge into the eye, which then sees an *upright virtual image*. The image distance *v* is equal to *D*, assuming that the lens is close to the eye.

Worked example

3 A converging lens of focal length 8.0 cm is used to magnify an object 2.0 mm tall.

a Where must the object be placed to form an image at the near point (v = 25 cm)?
b What is the height of the image?
c Is the image upright or inverted?

This question could be answered by drawing a ray diagram, but we will use the thin lens formula.

a $\frac{1}{f} = \frac{1}{v} + \frac{1}{u}$

$\frac{1}{8.0} = \frac{1}{-25} + \frac{1}{u}$ remembering that a virtual image must be given a negative image distance

u = 6.1 cm

b $m = -\frac{v}{u} = -\frac{-25}{6.1} = 4.1$

so the height of the image is 4.1 × 2.0 = 8.2 mm

c The magnification is positive, which means that the image is upright.

But the height of the image cannot be measured directly, so we are usually more concerned about the angular magnification, M, of a magnifying glass than its linear magnification, m.

$$M_{\text{near point}} = \frac{\text{angle subtended at the eye by the image formed at the near point}}{\text{angle subtended at the eye by the object placed at the near point}}$$

Looking at Figure 15.12:

$$M_{\text{near point}} = \frac{\theta_i}{\theta_o} = \frac{h_i/D}{h_o/D} = \frac{h_i}{h_o}$$

Note that this is numerically the same as the linear magnification, m (= $-v/u$), but because the height of a virtual image is not easily measurable we need to find an alternative method of calculating the magnification, and it is also desirable to be able to calculate the possible magnification directly from a knowledge of the focal length of the lens.

Looking at the similar triangles in Figure 15.12 containing the angle θ_i, we see that:

$$M_{\text{near point}} = \frac{h_i}{h_o} = \frac{D}{u}$$

But we want an equation that gives us the angular magnification in terms of f, not u. Multiplying the lens equation ($\frac{1}{f} = \frac{1}{v} + \frac{1}{u}$) throughout by v gives us:

$$\frac{v}{f} = \frac{v}{v} + \frac{v}{u}$$

Remember that in this situation $v = -D$ (the negative sign is included because the image is virtual), so we get:

$$-\frac{D}{f} = 1 - M_{\text{near point}}$$

or

$$M_{\text{near point}} = \frac{D}{f} + 1$$

This equation is given in the *Physics data booklet*.

Using the data from Worked example 3 gives $M_{\text{near point}} = \frac{25}{8} + 1 = 4.1$, as before.

Image at infinity

Forming an image at the near point provides the largest possible magnification, but the image can also be formed at infinity and this allows the eye to be more relaxed. Figure 15.13 shows that the object must be placed at the focal point.

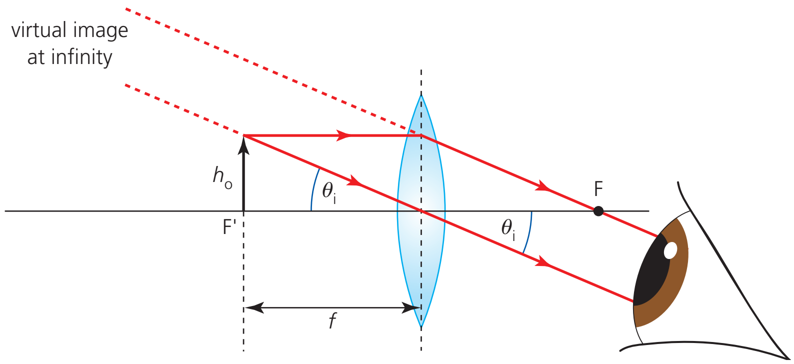


Figure 15.13 A simple magnifying glass with the image at infinity

From Figure 15.13 we see that $\theta_i = \frac{h_o}{f}$, so that:

$$M_{infinity} = \frac{\theta_i}{\theta_o} = \frac{\frac{h_o}{f}}{\frac{h_o}{D}}$$

$$M_{infinity} = \frac{D}{f}$$

This equation is given in the *Physics data booklet*.

By adjusting the distance between the object and the lens, the angular magnification can be adjusted from $\frac{D}{f}$ to $\frac{D}{f} + 1$, but the *lens aberrations* (see later) of high-power (small f) lenses limit the magnification possible with a single lens. A typical focal length for a magnifying glass is about 10 cm, which will produce an angular magnification between 2.5 and 3.5. More magnification would require a lens of greater curvature and too many aberrations.

20 a Draw an accurate ray diagram to show the formation of the image when an object is placed 5.0 cm away from a converging lens of focal length 8.0 cm.
b Use the diagram to determine the linear magnification.

21 Use the thin lens formula to predict the nature, position and linear magnification of the image formed by a converging lens of power +20 D when it is used to look at an object 4.0 cm from the lens.

22 What is the focal length of a converging lens that produces a virtual image of length 5.8 cm when viewing a spider of length 1.8 cm placed at a distance of 6.9 cm from the lens?

23 a Calculate the angular magnification produced by a converging lens of focal length 12 cm when observing an image at the near point.
b In what direction would the lens need to be moved in order for the image to be moved to infinity and for the eye to be more relaxed?
c When the lens is adjusted in this way, what happens to the angular magnification?

24 What power lens will produce an angular magnification of 3.0 of an image at infinity?

25 Two small objects that are 0.10 mm apart can just be distinguished as separate when they are placed at the near point. What is the closest they can be together and still be distinguished when a normal human eye views them using a simple magnifying glass that has a focal length of 8.0 cm?

26 a Where must an object be placed for a virtual image to be seen at the near point when using a lens of focal length 7.5 cm?
b Calculate the angular magnification in this position.

Predicting properties of virtual images formed by diverging lenses

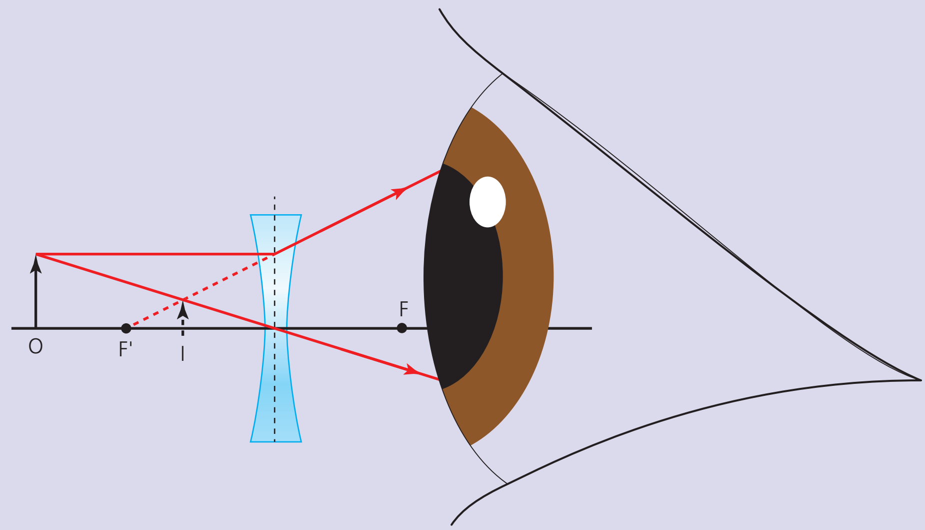
Because they do not form real images, diverging lenses have fewer uses than converging lenses. However, ray diagrams and the thin lens equation can be used for them in the same way as for converging lenses.

Worked example

4 A 2.0 cm tall object is placed 6.0 cm from a diverging lens of focal length 4.0 cm. Determine the properties of the image by:

a using a ray diagram
b using the lens equation.

a Figure 15.14 shows an image that is virtual, upright, 0.8 cm tall and 2.4 cm from the centre of the lens.



■ **Figure 15.14** Virtual upright image formed by a diverging lens (not to scale)

b $\frac{1}{f} = \frac{1}{v} + \frac{1}{u}$

$\frac{1}{-4.0} = \frac{1}{v} + \frac{1}{6.0}$

$v = \frac{-12}{5} = -2.4\,\text{cm}$

The negative sign represents a virtual image.

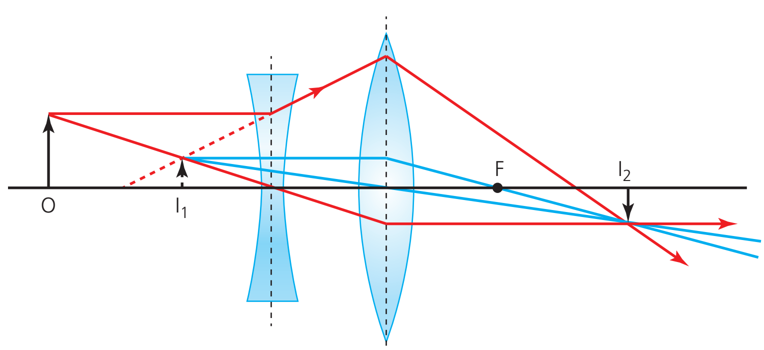
$m = \frac{-v}{u} = -\left(\frac{-2.4}{6.0}\right) = +0.40$; so the image size = 0.40 × 2.0 = 0.80 cm

The positive sign represents an upright image.

Combining lenses

If two lenses are used in an optical system, the final image can be predicted by treating the image formed by the first lens as the object for the second lens. Figure 15.15 shows an example in which the virtual image formed by the diverging lens in Figure 15.14 is used to form a second, real image by a converging lens of focal length 3.0 cm with its centre 3.1 cm from the centre of the diverging lens. The blue lines are just construction lines used to locate the top of the final image.

■ **Figure 15.15** Combining lenses (not to scale)



From a scale drawing we can see that the final image is real and inverted. It is located 6.6 cm from the converging lens and its size is 1.0 cm.

Alternatively, we can locate the image using the lens equation:

$$\frac{1}{3.0} = \frac{1}{v} + \frac{1}{(3.1 + 2.4)}$$

$$v = 6.6\,\text{cm}$$

The positive sign represents a real image.

$$m = \frac{-v}{u} = \frac{-6.6}{5.5} = -1.2$$

so:

final image size = 1.2 × 0.80 = 0.96 cm

The negative sign represents an inverted image.

Optical powers of lens combinations

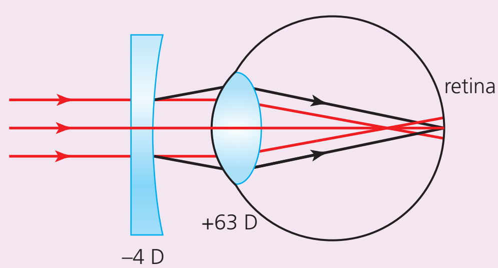
When two or more thin lenses are placed close together, the optical power of the combination is approximately equal to the sum of their individual powers. For example, combining a +4 D lens with a −1.5 D lens will produce a combined power of +2.5 D. In terms of focal lengths, combining a converging lens of focal length 25 cm with a diverging lens of focal length 67 cm will have a combined focal length of 40 cm.

Utilizations

Correcting vision defects

The distance between the lens and the retina in an adult human eye is typically about 1.7 cm. This means that a normal human eye has a focal length of about 1.7 cm when viewing a distant object (at the *far point*), which is equivalent to a power of about +60 D. The shape of the lens can be controlled so that the power can be varied in order to focus objects that are different distances away. For example, when observing an object 25 cm from the eye (at the standard *near point*) the focal length needs to be 1.5 cm, which is equivalent to a power of +67 D. In other words, the eye needs to *accommodate* objects at different distances by changing its power by up to +7 D.

Younger people can normally use the muscles in their eyes to change the power of their eyes by approximately +10D, but as people get older most of them gradually lose this ability, and by the age of 70 many are unable to achieve a wide range of focus. Most commonly, older eyes have insufficient optical power to be able focus on close objects and need spectacles with converging lenses to provide the extra power needed for reading.



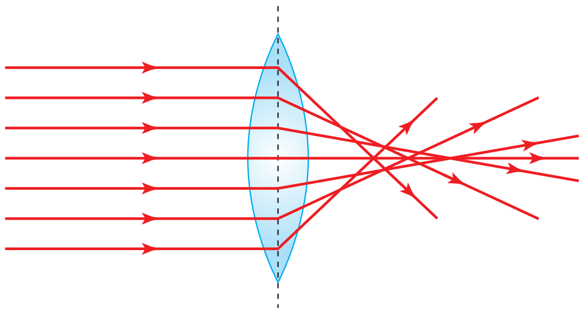
■ **Figure 15.16** Correcting *short-sight* (the red lines show the paths that the rays would follow without the lens)

Figure 15.16 shows a common eye defect in younger people. Light from a distant object is focused slightly in front of the retina. A simplified interpretation might be that the lens is 'too powerful' to form an image on the retina because it has a focal length of, for example, 1.6 cm instead of the required 1.7 cm (a power of +63 D instead of +59 D). This defect can be corrected by using spectacles of power −4 D (diverging lenses).

1 Find out how laser eye surgery can be used to correct vision defects and the circumstances under which it may be considered suitable or unsuitable.

■ Spherical and chromatic aberrations

Aberration is the term we use to describe the fact that, with real lenses, all the light coming from the same place on an object does not focus in exactly the same place on the image (as simple optics theory suggests). There are two principal kinds of aberration – spherical and chromatic.

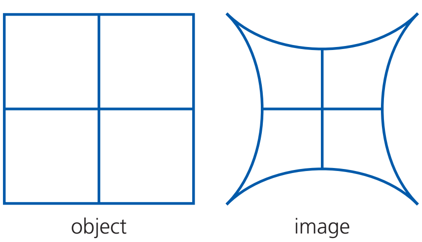


■ **Figure 15.17** Spherical aberration of monochromatic light (exaggerated)

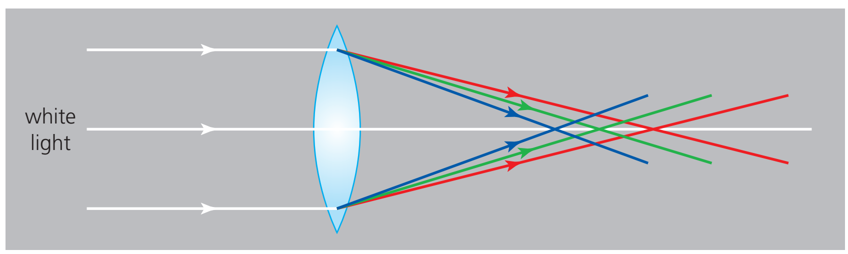
Figure 15.17 represents **spherical aberration**. This is the inability of a lens, which has surfaces that are spherically shaped, to focus parallel rays that strike the lens at different distances from the principal axis to the same point.

Spherical aberration results in unwanted blurring and distortion of images (see Figure 15.18), but in good-quality lenses the effect is reduced by adjusting the shape of the lens. However, this cannot completely remove aberration for all circumstances. The effects can also be reduced by only letting light rays strike close to the centre of the lens. In photography the size of the aperture (opening) through which light passes before it strikes the lens can be decreased to reduce the effects of spherical aberration. This is commonly known as 'stopping down' the lens, but it has the disadvantage of reducing the amount of light passing into the camera and may also produce unwanted diffraction effects.

Figure 15.19 represents **chromatic aberration**. Chromatic aberration is the inability of a lens to refract parallel rays of light of different colours (wavelengths) to the same focal point. Any transparent medium has slightly different refractive indices for light of different frequencies, so that white light may be dispersed into different colours when it is refracted. Typically, chromatic aberration leads to the blurring of images and gives images red or blue/violet edges.

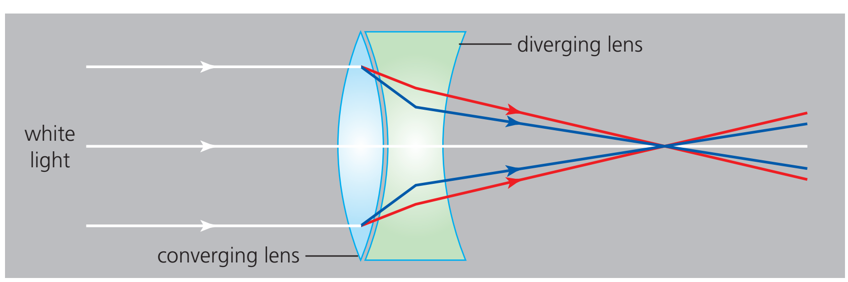


■ **Figure 15.18** Typical distortion produced by spherical aberration (exaggerated)



■ **Figure 15.19** Chromatic aberration

Chromatic aberration can be reduced by combining two or more lenses together. For example, a converging lens can be combined with a diverging lens (of a different refractive index), so that the second lens eliminates the chromatic aberration caused by the first (see Figure 15.20).



■ **Figure 15.20** Combining lenses of different refractive indices to correct for chromatic aberration

In the modern world we are surrounded by optical equipment capable of capturing video and still pictures and the quality of lenses has improved enormously in recent years. The quality of the images produced by the best modern camera lenses is highly impressive (Figure 15.21) and the improved detection of low light levels has meant that lenses (in mobile phones for example) can be very small, so that aberrations are less significant.

Figure 15.21 This lens achieves top-quality images by having a large number of lens elements

27 a What is the focal length of a diverging lens that will produce an image 8.0 cm from its centre when an object is placed 10.0 cm from the lens?

b List the properties of the image.

28 Suggest how the focal length of a diverging lens can be determined experimentally.

29 Two converging lenses with focal lengths of 10 cm and 20 cm are placed with their centres 30 cm apart. What is the linear magnification produced by this system when an object is placed 75 cm from the midpoint between the two lenses? Does this question have two different answers?

30 Make a copy of Figure 15.19 and show on it where a screen would have to be placed to obtain an image with blue edges.

31 Draw a diagram(s) to illustrate the improved focusing achieved by 'stopping down' a lens.

32 Suggest why lens aberrations tend to be worse for higher-power lenses.

33 In order to reduce chromatic aberration a converging lens of power +25 D was combined with a diverging lens of power −12 D. What is the focal length of this combination?

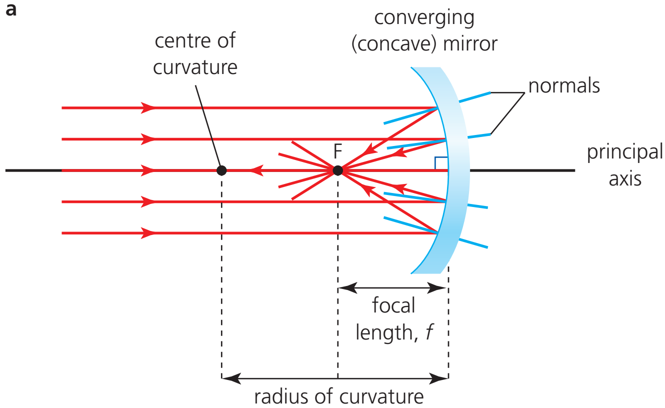
Converging and diverging mirrors

Mirrors with curved surfaces can also be used to focus images. The terminology and the principles involved are very similar to those already discussed concerning lenses. Figure 15.22 shows the action of spherically shaped reflecting surfaces on parallel wavefronts represented by rays. Once again, the theory will assume that *the rays are close to the principal axis* and *strike the mirror almost perpendicularly* (the diagrams are exaggerated for clarity).

The directions of the reflected rays can be predicted using the law of reflection (angle of incidence = angle of reflection). The concave surface (a) reflects the rays so that they converge to a real focal point, *F*, so the mirror is described as a *converging mirror*. The convex surface (b) reflects the rays so that they diverge from a virtual focal point, *F*, so the mirror is described as a *diverging mirror*. The distance from the **centre of curvature** of the spherical surface to the surface of the mirror is equal to twice the focal length, $2f$.

The properties of the image formed by a converging mirror can be investigated using an illuminated object and moving a screen (and/or the object) until a well-focused image is observed. Variations in the image can be observed when the mirror is moved, or if the mirror is exchanged for another with a different focal length.

a



b

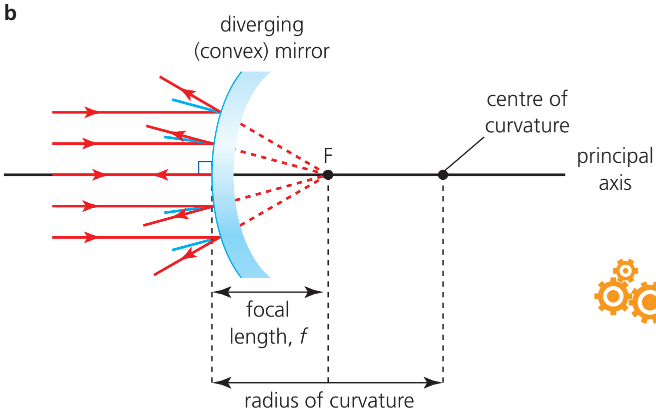


Figure 15.22 Reflection by spherical surfaces

Using ray diagrams to predict the properties of images in converging mirrors

As with lens diagrams, there are three rays, the paths of which we can always predict.

- An incident ray parallel to the principal axis will be reflected through the focal point, or be reflected so that it appears to come from the focal point.
- An incident ray passing through the focal point will be reflected parallel to the principal axis.
- An incident ray passing through (or directed towards) the centre of curvature will be reflected back along the same path.

Figure 15.23 uses these rays to predict the properties of images in a converging mirror.

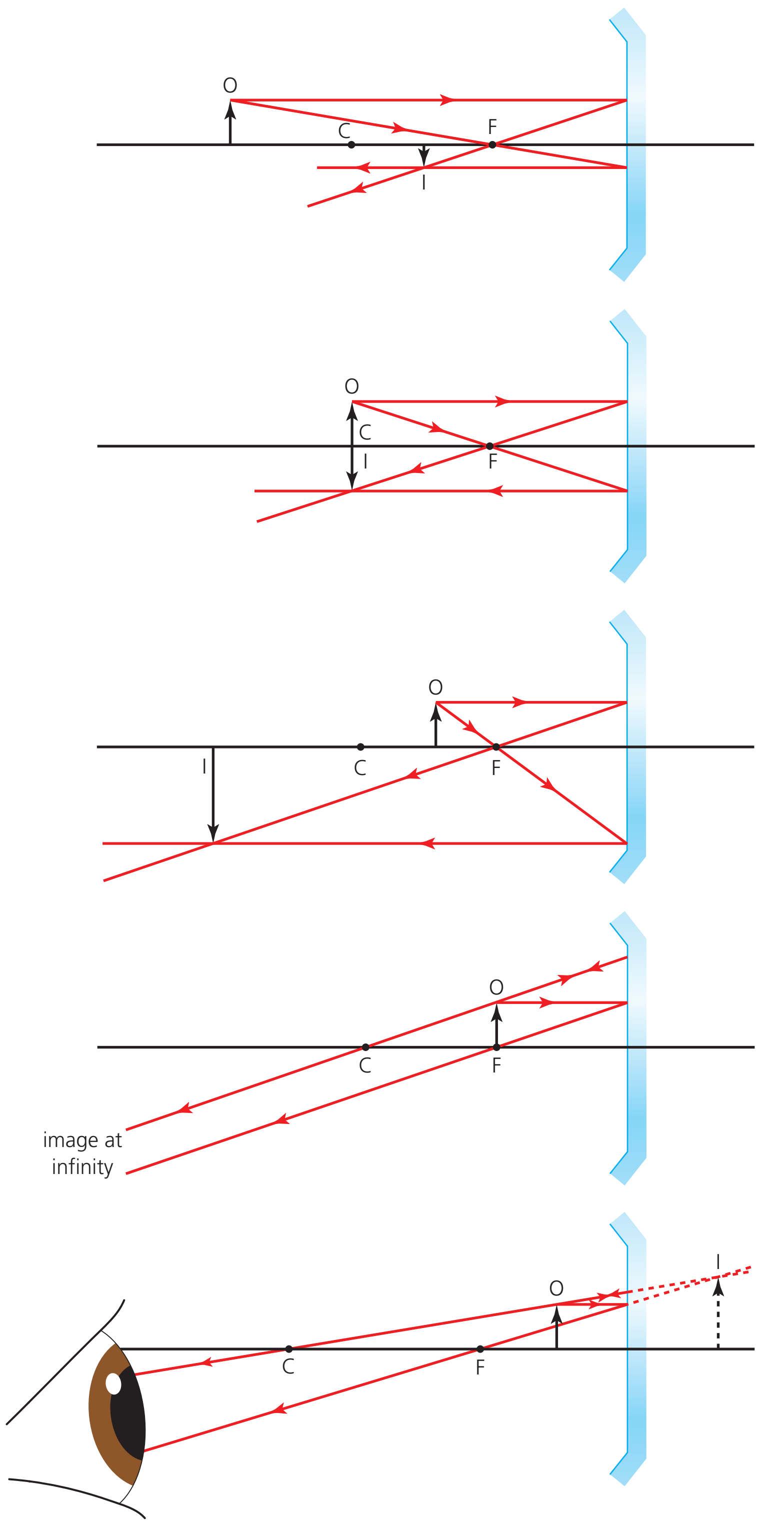


Figure 15.23 How the image changes as an object is brought closer to a converging mirror

Figure 15.23 shows us that as the object gets closer to the lens, the real inverted image gets larger and further away from the lens. But if the object is closer than the focal point the image is virtual, upright and magnified.

Linear magnification, $m = \frac{h_i}{h_o} = \frac{v}{u}$ (as with lenses)

Angular magnification, $M = \frac{\theta_i}{\theta_o}$ (as with lenses)

Worked example

5 When a 3.2 cm tall object is placed 5.1 cm from a converging mirror, a magnified virtual image is formed 9.7 cm from the mirror.

a What is the linear magnification?

b How tall is the image?

a $m = \frac{v}{u} = \frac{9.7}{5.1} = 1.9$

b $m = \frac{h_i}{h_o}$

$1.9 = \frac{h_i}{3.2}$

$h_i = 3.2 \times 1.9 = 6.1\,\text{cm}$

Diverging mirrors

Figure 15.24 show the formation of a diminished, upright, virtual image by a *diverging mirror*. This can be very useful when we need to see a wide field of vision, Figure 15.25 shows one application – a car's wing mirror.

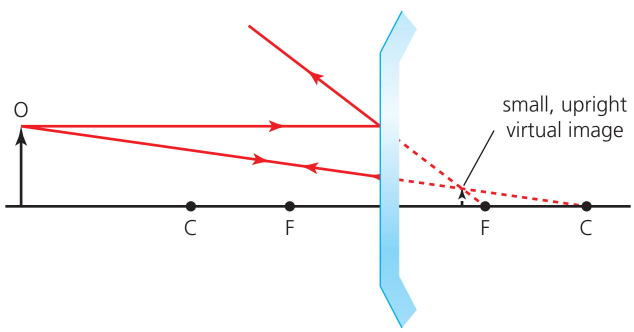


Figure 15.24 Producing small virtual images using a diverging mirror

Figure 15.25 Wide field of vision produced by a car's wing mirror

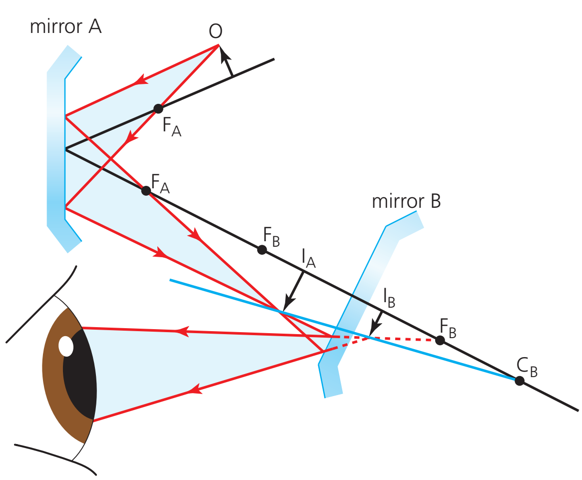


Figure 15.26 Forming an image using two curved mirrors

Mirror combinations

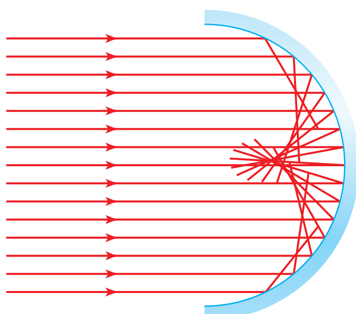
Ray diagrams for locating the image formed by two curved mirrors can be difficult to draw because they usually do not share the same single principal axis. Figure 15.26 shows an example. The object, O, forms a real, magnified, inverted image, I_A, after the light has been reflected by the converging mirror, A. The principal axis of A has been drawn in two positions, the second of which is also the principal axis of the diverging mirror, B. Remember that in ray tracing we always assume that the rays are close to the principal axis and strike the mirror perpendicularly, even though this may not be represented well in the diagrams.

I_A then provides the object that produces the (still) inverted, virtual image, I_B, when the rays reflect off the diverging mirror, B. (The blue line is a construction line used to determine the position of the top of the image.)

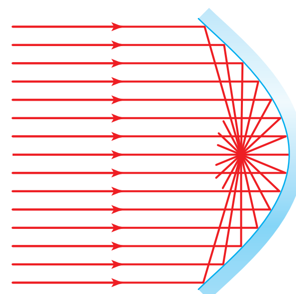
Spherical aberration in mirrors

We have been assuming that spherical surfaces produce perfect focuses, and that is an acceptable assumption for rays close to the principal axis striking the surface almost perpendicularly, but for many applications (especially for larger mirrors) we need to be more realistic. Figure 15.27 shows the effect of diverse reflections from a large spherical surface (the shape seen is often called a *caustic curve*).

Spherical aberration can be overcome by adapting the shape of the reflecting surface. **Parabolic reflectors** can produce much better focuses than spherical surfaces – see Figure 15.28. Receiving dishes for satellite broadcasts are a good example of this kind of converging reflector.



■ **Figure 15.27** Spherical aberration prevents a perfect focus



■ **Figure 15.28** A parabolic surface can produce a good focus

The same idea can be used in 'reverse'. If a point source of light (or other radiation) is placed at, or near, the focus of a parabolic reflector, the emerging beam will be parallel, or have a low divergence. The beams from a torch, car headlight or spotlight (Figure 15.29) are good examples of beams with only small divergence, which are produced by parabolic reflecting surfaces.

■ **Figure 15.29** The low divergence beam from a spotlight

34 Draw a ray diagram to determine the properties of the image formed when an object 1.5 cm tall is placed 7.0 cm from a converging mirror of focal length 5 cm.

35 a Draw a ray diagram of a diverging lens forming an image of an object placed between the mirror and its focal point.
b Describe the properties of the image.

36 a A make-up/shaving mirror uses a curved mirror. Describe the image seen.
b What kind of mirror is used, and typically how far away would a face be when using such a mirror?
c Suggest a suitable focal length for such a mirror.

37 Draw a ray diagram to locate the final image formed by the following optical arrangement. An object is placed 20 cm away from a large converging mirror of focal length 8 cm; the image formed is located 4 cm in front of a small converging mirror of focal length 5 cm. The two mirrors face each other.

38 Draw ray diagrams to represent:
a spherical aberration in a diverging mirror
b the production of the light beam from a car headlight.

15.2 (C2: Core) Imaging instrumentation *– optical microscopes and telescopes utilize similar physical properties of lenses and mirrors; analysis of the universe is performed both optically and by using radio telescopes to investigate different regions of the electromagnetic spectrum*

In this section we will look at how lenses and/or mirrors can be combined to produce optical images better than can be seen by the eye or by the use of a single lens. Similar ideas can then be applied to the use of other parts of the electromagnetic spectrum for imaging, in particular the use of radio waves in astronomy. Extending the range of human senses in these ways has contributed enormously to our expanding knowledge of both the microscopic world and the rest of the universe.

Optical compound microscopes

If we want to observe an image of a nearby object with a higher magnification than can be provided with a single lens, a second converging lens can be used to magnify the first image (see Figure 15.30). The lens closer to the object is called the **objective lens** and the second lens, closer to the eye, is called the **eyepiece lens**. Two lenses used in this way are described as a *compound microscope*. Note that the size of the lenses and their separation are *not* drawn to scale.

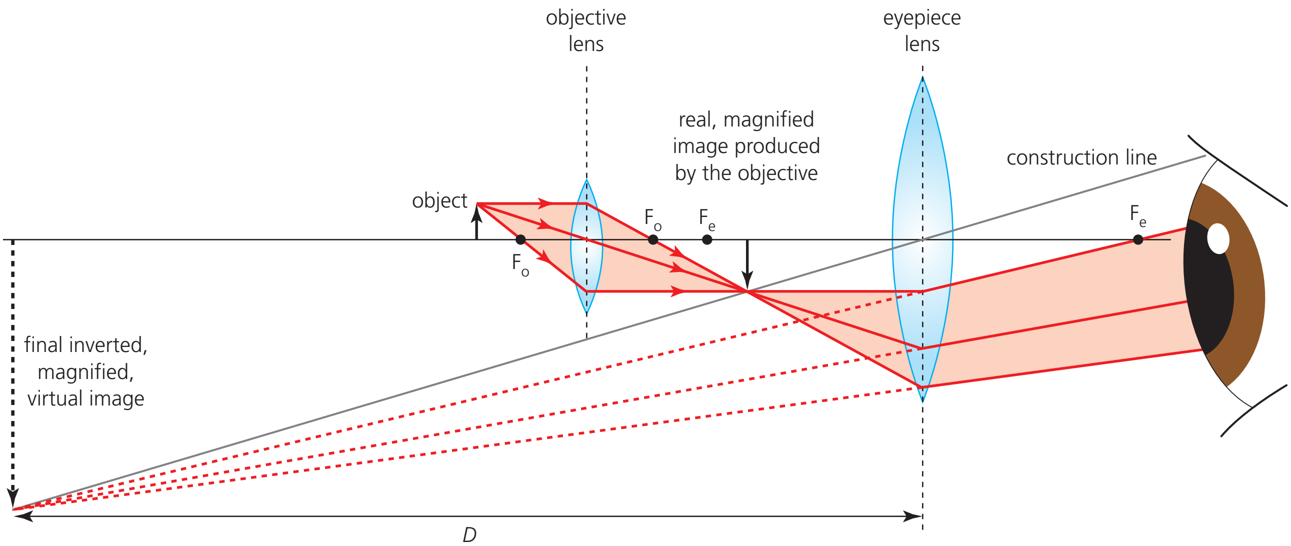


Figure 15.30 Compound microscope with the final image at the near point (normal adjustment)

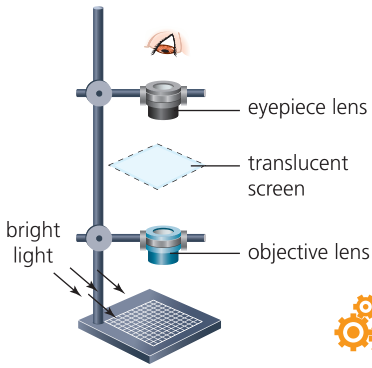


Figure 15.31 Investigating a model microscope

The object to be viewed under the microscope is placed just beyond the focal point of the objective lens, so that a real image is formed between the two lenses with a high magnification. The eyepiece lens is then used as a magnifying glass, and its position is adjusted to give as large an image as possible with the final virtual image usually at, or very close to, the near point of the eye. This is called *normal adjustment*.

To locate the image by drawing you need to find the point where the construction line through the centre of the eyepiece from the top of the first image meets the extension of the ray from the first image that passes through the focal point.

A model of a simple compound microscope can be investigated in a darkened room as shown in Figure 15.31. To begin with, a converging lens with a focal length of about 5 cm is used to form an inverted image of a brightly illuminated object (e.g. graph paper) on a **translucent** screen. Then, the position of a second, less-powerful lens is adjusted until a second (virtual) image of the first image is seen when looking through this eyepiece. The screen can then be removed and the two lenses used together to observe the scale, so that the magnification of the image can be estimated. As with many optical experiments, keeping the eye and all the components aligned is important for success.

Angular magnification

The angular magnification produced by a compound microscope is equal to the product of the linear magnification of the objective lens multiplied by the angular magnification of the eyepiece lens. For an image at the near point:

$$M_{\text{overall}} = m_{\text{objective}} \times M_{\text{eyepiece}} = \left(\frac{-v}{u}\right)_{\text{objective}} \times \left(\frac{D}{f} + 1\right)_{\text{eyepiece}}$$

This equation is *not* given in the *Physics data booklet*. If the final image is at infinity (for less eye strain) the +1 term can be omitted.

Worked example

6 A compound microscope contains an objective lens of focal length 0.80 cm and an eyepiece lens of focal length 5.4 cm. The microscope is adjusted to form a final image at the near point of the eye for an object placed 0.92 cm in front of the objective.

a Determine the distance between the two lenses.

b What is the angular magnification of the image?

a First find the image distance for the objective:

$$\frac{1}{f} = \frac{1}{v} + \frac{1}{u}$$

$$\frac{1}{0.80} = \frac{1}{v} + \frac{1}{0.92}$$

$v = 6.1$ cm

Then find the object distance for the eyepiece:

$$\frac{1}{5.4} = \frac{1}{-25} + \frac{1}{u}$$

(remembering that virtual image distances are negative)

$u = 4.4$ cm

distance between lenses = $v + u = 6.1 + 4.4 = 10.5$ cm

b $$M = \left(\frac{v}{u}\right)_{\text{objective}} \times \left(\frac{D}{f} + 1\right)_{\text{eyepiece}}$$

$$M = \left(\frac{6.1}{0.92}\right) \times \left(\frac{25}{5.4} + 1\right) = 37$$

The exact angular magnification of a microscope clearly depends on where the object and final image are located, but an approximate guide to the angular magnification of a compound microscope can be obtained from the focal lengths and the distance between the lenses, L: $M \approx \frac{DL}{f_o f_e}$. (This equation predicts $M \approx 45$ for Worked example 6.) This confirms that shorter focal lengths will produce higher magnifications; but, as with magnifying glasses, the higher curvatures associated with higher of shorter focal lengths will introduce lens aberrations and reduce the quality of the images produced.

Resolution

Although the magnification produced by a microscope is obviously important, *resolution* is usually of more significance in optical instruments. In general, high resolution can be described as the ability to see detail in an image. Two images that can be seen as separate are said to be *resolved*. To understand the difference, consider a picture on a phone or computer screen – it can be very easy to magnify, but that often results in a poorer-quality image.

A high-quality microscope, or telescope, will produce images with good magnification *and* resolution; a poorer-quality instrument may produce high magnification but the resolution will be disappointing.

The *magnification* of an optical system is largely dependent on the focal lengths of the lenses, but the *resolution* of a system depends on the quality of the lenses, the diameter of the objective and the wavelength of the radiation being detected. High overall resolution also assumes that the properties of the surface detecting the waves, for example the separation of pixels in a camera or the separation of cells on the retina of the eye, will not have an adverse effect.

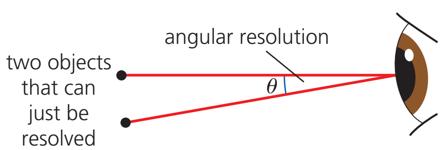


Figure 15.32 Angular resolution of the eye

A normal human eye can see two similar objects placed at the near point as separate if they are approximately at least 0.1 mm apart. Resolution is best represented by the angle subtended by these points, $\theta = \frac{0.1}{250} \approx 4 \times 10^{-4}$ rad (see Figure 15.32). Better resolution is represented by a *smaller* angle between two points that can just be seen as separate.

Using a good microscope with an angular magnification of, say, 50 could improve the resolution by the same factor, so that it would then be possible to separate two points subtending an angle of $\frac{(4 \times 10^{-4})}{50} = 8 \times 10^{-6}$ rad, which corresponds to a linear separation of 2×10^{-3} mm at the near point.

Rayleigh's criterion

The diffraction of waves as they pass into the eye, or an optical instrument, is the main factor limiting resolution, and the amount of diffraction depends on the wavelength, λ, and the width of the aperture, b (Chapter 4). *Rayleigh's criterion* is a guide to resolution for waves passing through circular apertures (the theory was discussed in Chapter 9, but is not needed here):

> Two objects are considered to be resolvable if the angle, θ, that they subtend at the eye or optical instrument is bigger than $1.22\lambda/b$.

Worked example

7 Use Rayleigh's criterion to estimate the angular resolution of the human eye.

Assuming the average wavelength of light is 6×10^{-7} m and the diameter of the pupil (in bright light) is 2 mm:

$$\theta = \frac{1.22\lambda}{b} = 1.22 \times \frac{6 \times 10^{-7}}{2 \times 10^{-3}} \approx 4 \times 10^{-4} \text{ rad}$$

Which is in reasonable agreement with actual observations.

Applying Rayleigh's criterion to the resolution achieved by optical instruments, we can see that resolution would be improved by using shorter wavelengths and wider apertures. Using wider apertures has the added advantage of admitting more light and producing brighter images, although larger lenses may have aberration problems. Using light of smaller wavelengths (the blue/violet end of the spectrum) can improve resolution but, of course, any coloured effects would be lost.

Rayleigh's criterion can be used as a guide to resolution, but there are other factors involved – for example placing an oil with a high refractive index between the specimen and the eyepiece can improve resolution.

Utilizations

Electron microscopes

The resolution of a microscope will be improved if radiation with a shorter wavelength than light can be used to examine the object. Waves of the electromagnetic spectrum with shorter wavelengths than light are ultraviolet, X-rays and gamma rays, but none of these are as easily produced, controlled and detected as a beam of electrons.

Like all particles, electrons have wave properties but, because their mass is so small, electron wave properties are relatively easily observed; electrons in a beam have typical wavelengths of about 10^{-10} m. This is about 5000 times smaller than the average wavelength of visible light, so that resolution can be improved by the same factor by using a beam of electrons rather than a beam of light photons.

■ **Figure 15.33**

Beams of electrons can be produced by accelerating potential differences of a few thousand volts and their wavelength can be adjusted by changing the p.d. Because electrons are charged, they can be focused by electric or magnetic fields.

Of course, electrons cannot be 'seen' using our eyes, so their energy needs to be converted to light to form a visible image (see Figure 15.33).

1 Figure 15.33 shows a living bed-bug. Use the internet to find out if placing such an organism in an electron beam is harmful.

39 An object is placed 5.0 mm from the objective lens of a two-lens compound microscope. The eyepiece of the microscope has a focal length of 4.0 cm.
  a If the linear magnification produced by the objective lens is 5.0, what is its focal length?
  b What is the overall angular magnification of the microscope when observing an image of this object at infinity?

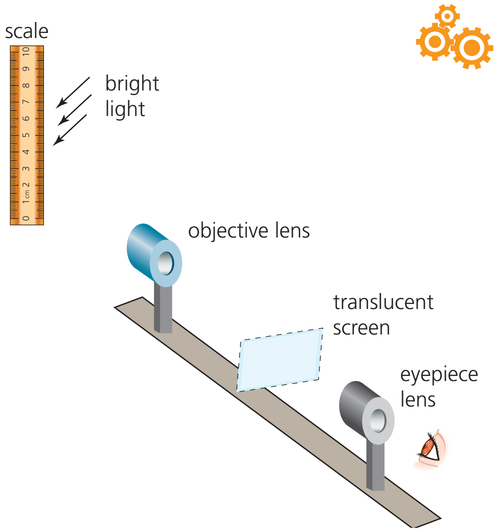
40 **a** If the diameter of the objective lens in a microscope is 1.2 cm, estimate its angular resolution assuming an average wavelength of light of 5.5×10^{-7} m.
  b If the microscope produces an angular magnification of 80, estimate the minimum separation of two points that can just be resolved by a normal human eye.
  c What assumption did you make in answering (b)?

41 Suggest why placing a transparent oil between a specimen and an objective lens can improve the resolution of a microscope.

■ Simple optical refracting telescopes

A telescope is an optical instrument designed to produce an angular magnification of a distant object. Images are formed by the processes of refraction or reflection. (Reflecting telescopes will be considered in the next section.)

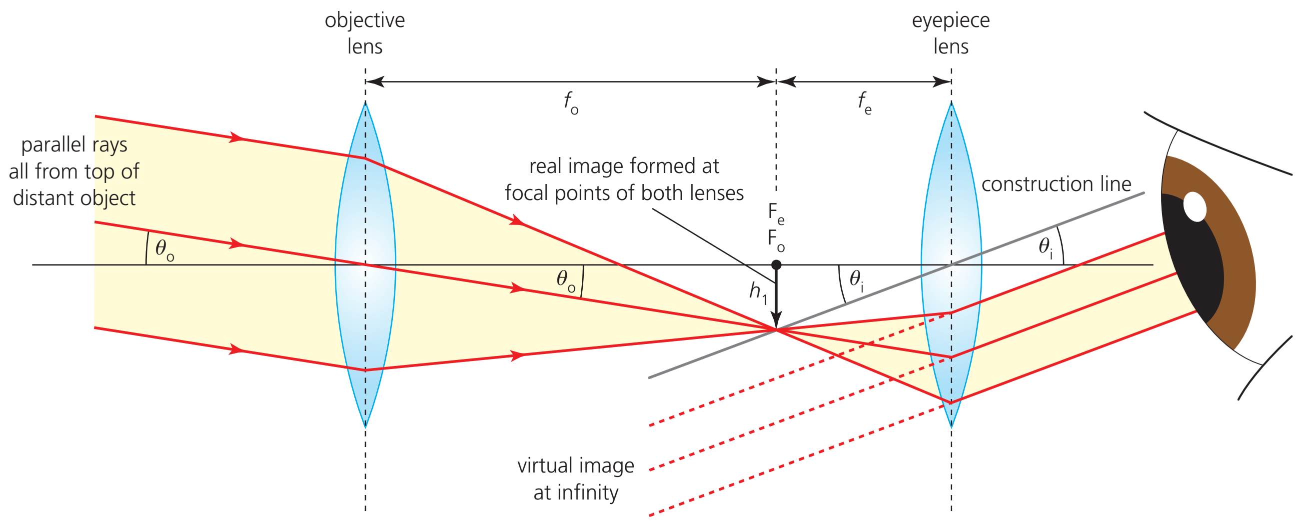
In a simple refracting astronomical telescope there are two lenses that together produce an *inverted*, virtual image. Such telescopes are usually of little use for looking at objects on Earth; this is why they are sometimes described as 'astronomical' – for use in astronomy. Light rays arriving at the objective lens of a telescope can be considered to be parallel because the source of light is such a long distance away. Consequently a small, inverted real image will be formed at the focal point of the objective lens. A second lens, the eyepiece, is then used as a magnifying glass to magnify this image.



■ **Figure 15.34** Investigating a model telescope

A model of an astronomical telescope can be investigated in a darkened room as shown in Figure 15.34. To begin with, a converging lens with a focal length of about 50 cm is used to form an inverted image of a brightly illuminated object (e.g. a scale) on a translucent screen. Then the position of a second, more powerful lens is adjusted until a second virtual image of the first image is seen when looking through the eyepiece. The screen can then be removed and the two lenses used together to observe the scale, and the magnification of the image can then be estimated. As with many optical experiments, keeping the eye and all the components aligned is important for success.

Figure 15.35 shows the basic construction of a two-lens astronomical telescope. The telescope is usually adjusted so that the final image is at infinity so that the eye can be relaxed when observing it for extended periods of time. This is called using the telescope in 'normal adjustment' and the image formed by the objective must be formed at the focal point of the eyepiece. When adjusted in this way, the distance between the lenses is the sum of their focal lengths. The direction to the top of the final image is located by drawing a construction line through the centre of the eyepiece from the top of the first image.



■ **Figure 15.35** Simple refracting telescope with the final image at infinity (normal adjustment)

The angular magnification (in this adjustment) can be determined by examining the two triangles involving h_1:

$$M = \frac{\theta_i}{\theta_o} = \frac{\frac{h_1}{f_e}}{\frac{h_1}{f_o}}$$

$$M = \frac{f_o}{f_e}$$

This equation is given in the *Physics data booklet*.

Clearly, a higher magnification is obtained by using an objective lens with a longer focal length (lower power) and an eyepiece lens with a smaller focal length (higher power). But lens aberrations of high-power eyepieces limit the angular magnification possible.

If a telescope, or binoculars, are required to produce an upright image (for non-astronomical use) then another lens, or prism, must be added to the simple design shown in Figure 15.35 to invert the image.

The quality and diameter of the objective lens are the most important factors when considering the quality of the image in any kind of optical instrument. A larger objective has two advantages:

- Most importantly, it collects more light to produce a brighter image (to see dimmer and more distant or smaller objects).
- There will be less diffraction, which improves the resolution of images.

However, larger objectives also have aberration problems, which inevitably reduce the quality of the images.

42 A student hopes to construct an astronomical refracting telescope that produces an angular magnification of at least 100.
  a If she chooses an objective lens of focal length 68 cm, what is the minimum power needed for the eyepiece?
  b Explain why this telescope may not produce high-quality images.

43 Venus has a diameter of about 12 000 km.
  a What angle does it subtend at the eye when it is a distance of 2.0×10^8 km from Earth? (Assume that all of Venus is visible.)
  b A refracting telescope with an objective lens of focal length 120 cm is used to observe Venus. What is the diameter of the image formed by this lens?
  c An eyepiece lens of focal length 1.5 cm is used to form a final virtual image of Venus at infinity. What is the angle subtended by the image at the eyepiece?
  d Use your answers to (a) and (c) to confirm that the overall angular magnification of the telescope is given by $M = f_o/f_e$.

44 A refracting telescope consists of lenses of focal length 86 cm and 2.1 cm.
  a Which lens is the eyepiece?
  b Calculate the angular magnification in normal adjustment.
  c If the objective lens was replaced with another having twice the diameter but the same focal length, how would the image change?

45 Suggest how the use of a third lens in a refracting telescope can result in an upright image.

ToK Link

Can we trust our own senses?

However advanced the technology, microscopes and telescopes always involve sense perception.Can technology be used effectively to extend or correct our senses?

Our eyes collect and focus light, and then electrical signals are sent along the optic nerve to our brains. The brain processes the information and the result is what we call an image, and we would describe this as real ('seeing is believing'). But our ability to observe is well known to be fallible, and simple optical illusions demonstrate how easily we can be fooled. In Figure 15.36 our eyes tell us that squares A and B are different shades, but scientific measurement would correctly inform us that they are the same.

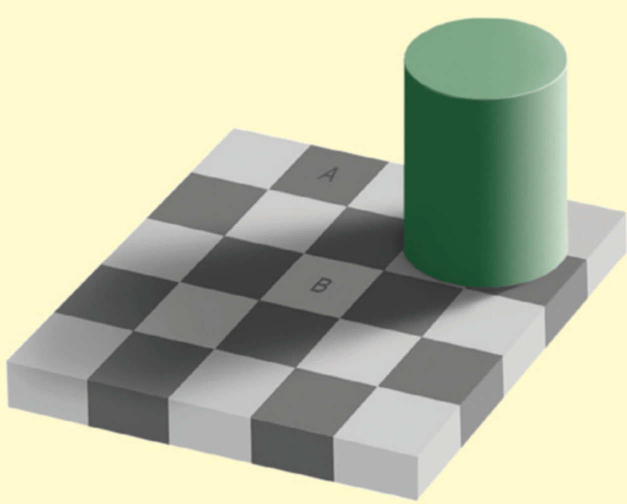


Figure 15.36 Squares A and B are exactly the same shade of grey!

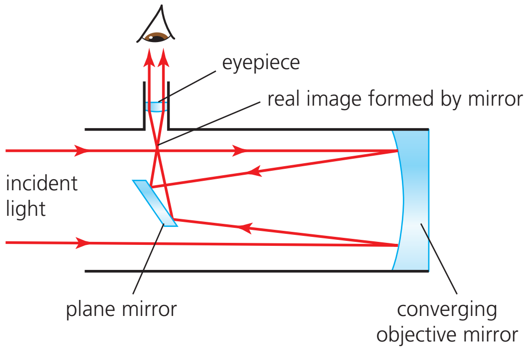


Figure 15.37 A reflecting telescope with a Newtonian mounting

Simple optical reflecting telescopes

A reflecting telescope has an objective mirror (converging) rather than a refracting objective lens. The image formed is observed through an eyepiece lens. Figure 15.37 shows the basic design of a reflecting telescope first described by Isaac Newton in 1668. It is described as having a **Newtonian mounting**. A small plane mirror has to be positioned in the incident beam in order to reflect the light into the eyepiece, which is positioned to the side of the main body of the telescope. Without this arrangement the observer would need to place their head in the incident beam. Of course it has the disadvantage that the plane mirror will prevent some of the light in the incident beam from reaching the converging mirror.

Worked example

8 Two stars separated by a distance, s_o, subtend an angle, θ_o, of 5.3×10^{-5} rad when viewed from Earth. The stars are observed through a Newtonian telescope with a converging mirror of focal length 3.4 m.

a Calculate the separation, s_i, of these two stars in the image formed by the converging mirror. (Assume they are the same distance from Earth.)

b If the eyepiece has a focal length, f_e, of 4.5 cm and is used to form a final image at infinity, what is the overall angular magnification produced by the telescope?

a

$$m = \frac{s_i}{s_o} = \frac{v}{u}$$

The incident light rays are (almost) parallel, so the image is formed at the focal point, and therefore $v = f$.

$$s_i = \frac{s_o}{u} \times f$$

But $\theta_o = \frac{s_o}{u}$, so:

$$s_i = \theta_o \times f = 5.3 \times 10^{-5} \times 3.4$$
$$= 1.8 \times 10^{-4}\,\text{m}$$

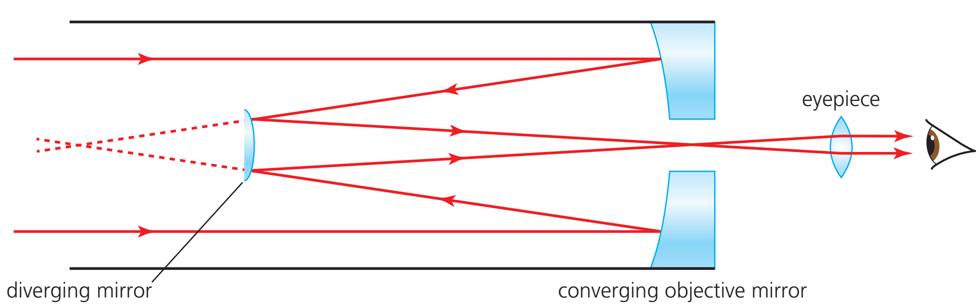
b The image from the mirror must be formed at the focal point of the eyepiece lens for the final image to be at infinity.

$$\text{angle subtended by image, } \theta_i = \frac{s_i}{f_e} = \frac{1.8 \times 10^{-4}}{0.045}$$
$$= 4.0 \times 10{-3}\,\text{rad}$$

$$M = \frac{\theta_i}{\theta_o} = \frac{4.0 \times 10^{-3}}{5.3 \times 10^{-5}} = 75$$

This could be found directly from $M = \frac{f_o}{f_e}$.

Figure 15.38 A reflecting telescope with a Cassegrain mounting



The **Cassegrain mounting** (designed by Laurent Cassegrain in 1672) is an alternative design that reflects the rays off a second (diverging) mirror back through a hole in the objective mirror – see Figure 15.38. This arrangement enables the user to look in the same direction as that from which the light is coming. Using a diverging mirror in this way produces extra magnification in a compact design.

Worked example

9 Consider Figure 15.38. The rays converging to the small diverging mirror would otherwise have formed an image 24 cm behind the mirror, but have been reflected to a focus 1.36 m away near the hole in the converging objective mirror. What extra magnification has this introduced compared to the use of a plane mirror in a Newtonian mounting with an objective mirror and eyepiece of similar focal lengths?

$$m = \frac{v}{u} = \frac{1.36}{0.24} = 5.6$$

Reflecting telescopes are still popular today and they have some important advantages over refracting telescopes, including:

- The light does not have to pass through a refracting medium, so there is no chromatic aberration.
- The light does not have to pass through a refracting medium, so there is no absorption or scattering.
- The objective has only one active surface so high-quality, larger-diameter objectives of the right shape are easier and cheaper to produce.

For these reasons the majority of optical astronomical telescopes used for research are reflectors. They are also popular with amateur astronomers. Figure 15.39 show the supporting structure of the world's largest reflecting telescope, Gran Telescopio Canarias. Figure 15.40 shows a smaller reflecting telescope for individual use.

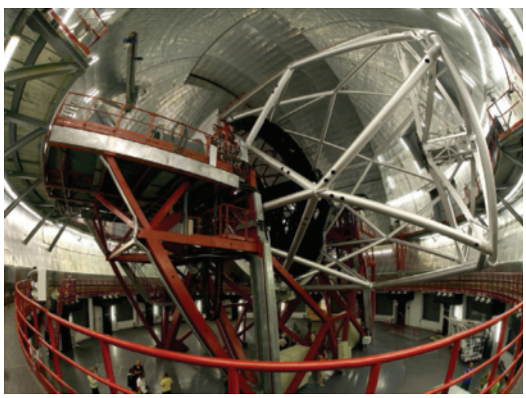


Figure 15.39 Reflecting telescope in the Canary Islands, Spain

Figure 15.40 Smaller reflecting telescope for amateur use

46 Determine the angular magnification of a Newtonian reflecting telescope that has a converging mirror of focal length 6.7 m and an eyepiece lens of focal length 1.8 cm. Assume that the final image is at infinity.

47 Increasing resolution and light-gathering ability is achieved by using larger mirrors. Explain why telescopes cannot be improved by simply making them bigger and bigger.

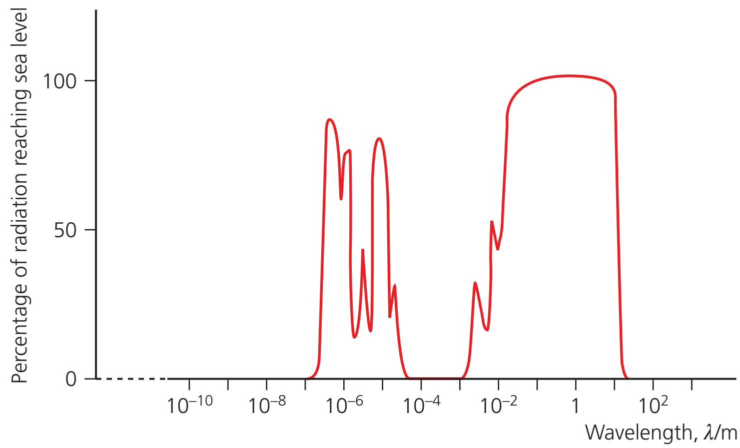
48 Use the internet to find out some of the reasons why an amateur astronomer would choose one of the two basic types of reflecting telescope described in this section (rather than the other).

Satellite-borne telescopes

Terrestrial telescopes (those on the Earth's surface) receive waves that have passed through the Earth's atmosphere. However, the atmosphere reflects, scatters, refracts and absorbs some of the incoming radiation, and these effects can significantly affect the quality of images formed by terrestrial telescopes. Examples of the effects of the atmosphere on radiation include stars viewed from the Earth's surface 'twinkling' because of the constantly changing effects of refraction, and clear skies appearing blue because shorter wavelengths of light scatter more from the atmosphere than longer wavelengths.

Optical astronomical telescopes also have particular problems – they can only be used at night, they can only be used if there are no clouds, and they are affected by light from the Earth being scattered by the atmosphere after dark ('light pollution').

The effects of the atmosphere are very dependent on the wavelengths of the radiation involved, as shown in Figure 15.41.



■ **Figure 15.41** How the Earth's atmosphere affects incoming radiation

There are a number of outstanding features in Figure 15.41:

- Radio waves, (e.g. with a typical wavelength of about 1 m) and microwaves are almost unaffected by passing through the atmosphere.
- Infrared is strongly absorbed (Chapter 8).
- Ultraviolet and X-ray radiations are mostly absorbed by the atmosphere.

An obvious way of reducing the effects of the atmosphere is to place telescopes in locations that are at high altitudes, with good weather conditions and which are far from towns and cities. See Figure 15.42.

With the considerable improvements in satellite technology in recent years, it has become possible to place significant numbers of space telescopes on satellites ('satellite-borne') in orbit around the Earth. This has produced an enormous amount of data (collected and analysed using high-power computers), new discoveries of less intense or more distant sources, or those emitting different kinds of radiation, and impressive high-resolution images.

■ **Figure 15.42** Telescopes at Mauna Kea, Hawaii

The Hubble telescope is probably the most well-known orbiting telescope, with many of its spectacular images well known and freely available around the world. The telescope was launched in 1990 and was named after the famous American astronomer Edwin Hubble. It has a mass of about 11 tonnes and orbits approximately 560 km above the Earth's surface, taking 96 minutes for one complete orbit. One of the greatest achievements of astronomers using the Hubble telescope has been accurately determining the distances to very distant stars, enabling a much improved estimate for the age of the universe.

Because such projects are expensive and the data obtained of interest to astronomers worldwide, they are often joint ventures between countries.

Utilizations

A different kind of observatory

Before the invention of the telescope, astronomers throughout the world made impressively accurate observations with their unaided eyes and by using a range of different devices to measure small angles.

More than 100 years after the discovery of the telescope, between 1727 and 1734, Maharaja Jai Singh II built an impressive observatory at Jaipur in India, which consisted of 14 large geometrical structures to assist astronomy with the unaided eye (Figure 15.43). The biggest of these is 27 m tall and is the largest sundial in the world. Its shadow can be seen to move at a rate of up to 6 cm every minute.

■ **Figure 15.43** Jantar Mantar at Jaipur, India

The purpose of the structures was to measure time and the apparent motions of the planets and stars, but also to be impressive structures in themselves and to stimulate interest in the newly developing science of astronomy. In India at that time, astronomy and astrology were closely connected, as they had been throughout the world in nearly all civilizations (and even today for many people).

1 Many people believe that the positions of the Moon, stars and planets can influence our individual lives and our futures. Do you think that this is possible? Explain your answer.

Non-optical telescopes

Nature of Science

Technological advances in astronomy

The invention of the optical telescope happened over 500 years ago; there is no general agreement about who was responsible although a German spectacle maker, Hans Lippershay, is often credited. Certainly Galileo adapted and improved early designs and his observations of moons orbiting Jupiter are well known. This was presented as evidence that the Earth may not be the centre of the universe and is an early example of the dramatic advances in human knowledge that can achieved by using instruments to extend our observations.

For most of the subsequent 500 years, astronomy has relied on the detection of visible light to provide information, but radiation from all other parts of the electromagnetic spectrum also arrive at the top of the Earth's atmosphere from space. Telescopes and sensors capable of detecting infrared, ultraviolet and X-rays have now been placed on orbiting satellites, and the data obtained leads to knowledge about the universe that can be very different from that obtained from light alone. For example, new sources of radiation have been discovered (for example gamma ray bursts and X-ray binary stars) and our knowledge of how the universe began has improved considerably.

Radio astronomy, in particular, is a highly advanced technology.

Radio telescopes

Radio waves are emitted by a wide variety of sources in space and they are mostly unaffected by passing through the Earth's atmosphere (see Figure 15.41), so radio telescopes can be terrestrial and, unlike visible light telescopes, they can be used 24 hours a day.

Some sources have been discovered from their radio wave emissions because they do not emit significant visible light, but radio waves are also emitted as part of the spectrum of elements (for example hydrogen, the most common element in the universe, emits a significant radio wavelength of 21 cm). In this way radio astronomy has helped to map the universe.

Single dish radio telescopes

A single 'dish' radio telescope uses a reflector (usually parabolic) to focus the radio waves to a detector (**aerial**) placed at the focus. Figure 15.44 shows the Parkes radio telescope in Australia, which has a diameter of 64 m.

Figure 15.44 The Parkes telescope in New South Wales, Australia

We know that a guide to the angular resolution of a telescope may be determined from $\frac{1.22\lambda}{b}$, where b is the width of the aperture/dish. The wavelengths to be investigated are predetermined by the nature of the investigation, and because radio wavelengths are much longer than light wavelengths, good resolution is much more difficult to achieve.

The most significant factor controlling the resolution of a single-dish radio telescope is the width of the dish. Larger dishes will produce higher resolution but, unfortunately, larger dishes are also much more expensive; it is also more difficult to maintain their precise shape and more difficult to steer them to the desired direction. It should also be stressed that larger dishes will collect more energy, so that more distant and dimmer sources can be detected. The largest single-dish radio telescope in the world is at the Arecibo Observatory in Puerto Rico. It has a diameter of 305 m – this is only possible because the surrounding landscape has been used to help support the structure.

Worked example

10 a Determine the resolution of the Parkes telescope (Figure 15.44) if it was used to detect the 21 cm wavelength of hydrogen.

b Would this telescope be able to resolve two stars emitting radio waves that were 2.3×10^{13} km apart and (both) 6.7×10^{15} km from Earth?

a $\frac{1.22\lambda}{b} = 1.22 \times \frac{0.21}{64}$

$= 4.0 \times 10^{-3}$

This is worse than optical telescopes or the human eye.

b angle subtended at Earth by the stars $= \frac{2.3 \times 10^{13}}{6.7 \times 10^{15}} = 3.4 \times 10^{-3}$

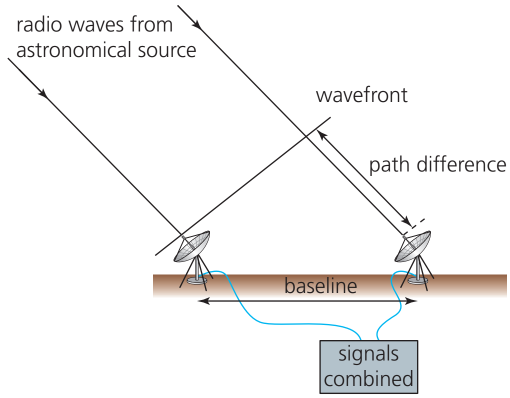
This angle is less than 4.0×10^{-3}, so the stars will *not* be resolvable.

Radio interferometry telescopes

There is an alternative method of improving the resolution of astronomical telescopes other than by constructing larger dishes. If the signals from an arrangement of two or more synchronized smaller telescopes can be combined (electronically), then the overall angular resolution will be approximately equal to that of a single dish with a diameter equal to the separation of the individual dishes. When using the equation for angular resolution, $\theta = \frac{1.22\lambda}{b}$, b becomes the separation of the telescopes rather than the width of each one.

There will be a *path difference* between a wavefront emitted from an astronomical source arriving at different telescopes (Figure 15.45). When the signals are combined, an interference pattern (Chapters 4 and 9) will be produced, the spacing and centre of which can be used to accurately determine the direction to the source of radiation.

Figure 15.45 Radio interferometry telescopes



Resolution is improved by using more telescopes in a regular pattern (compare with the use of diffraction gratings, discussed in Chapter 9) and there are a number of different ways in which a pattern of telescopes might be arranged, however the details are not needed for this course. Receiving enough energy from distant and dim sources is always an important issue, so the total combined collecting area of the telescopes must still be as large as possible (which is another reason why more dishes are preferable).

The individual telescopes may be arranged in an **array** relatively close together, see Figure 15.46, or they may be separated by a long distance (a *long baseline*), which can be as much as thousands of kilometres long and in different countries, although longer distances introduce technological problems.

Figure 15.46 An interferometer array

Worked example

11 Compare the theoretical resolution of two radio telescopes of dish diameter 50 m separated by a distance of 1 km with one of the telescopes used individually.

The effective aperture has been increased by a factor of 1000/50 = 20. So the angular resolution of the two together is 20 times smaller (better). The combination will also have double the receiving area and will therefore be able to detect less-intense sources.

49 Suggest some reasons why the 'weightless' and airless environment of a satellite orbit has advantages for reflecting telescopes.

50 Make a list of the advantages of placing telescopes on orbiting satellites.

51 When detecting radio waves of frequency 1420 MHz, what is the minimum diameter of a single-dish reflecting telescope needed to resolve two stars that are 4.7×10^{16} m apart and 1.5×10^{19} m from Earth?

52 What angular resolution would be obtained from an interferometer using two radio telescopes 540 m apart using radio waves of wavelength 0.18 m?

53 An angular resolution of 1 arcsecond is obtained using interferometry techniques. What separation of two telescopes would be needed to achieve this value when receiving radio waves with a frequency of 6 GHz?

54 Use the internet to research the latest developments in the SKA project.

55 Suggest what problems may arise when using interferometry techniques between telescopes that are thousands of kilometres apart.

15.3 (C3: Core) Fibre optics *– total internal reflection allows light or infrared radiation to travel along a transparent fibre; however, the performance of a fibre can be degraded by dispersion and attenuation effects*

In Chapter 4 we introduced the ideas of *critical angle* and *total internal reflection*, and briefly discussed how these could be used in, for example, optic fibre endoscopes to obtain images from inside the human body. In this section we will discuss optic fibres in more detail, with respect to their advantages in the **transmission** of data.

Optic fibres, wires and wireless communication

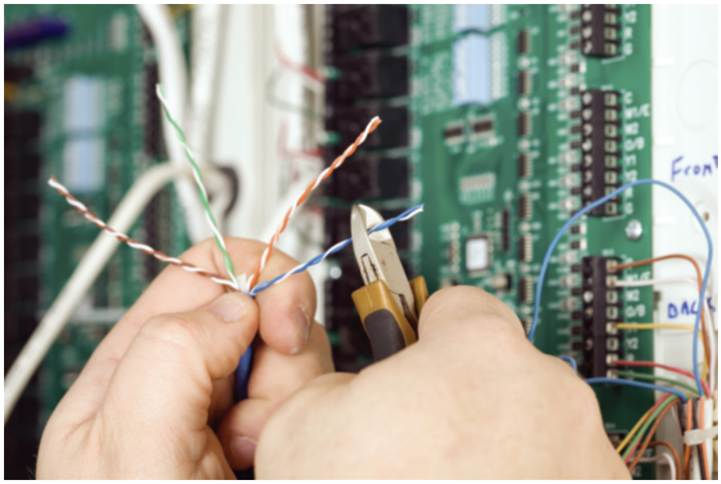
If we want to communicate over long distances and/or quickly transmit a large amount of data, we have two choices: (i) using electromagnetic waves *through the air*, (ii) sending signals *along some kind of cable* (insulated wire(s) or fibre(s)). There are many factors affecting the choice between these, and the advantages and disadvantages vary with the particular circumstances. But the majority of data transfer worldwide is still transmitted using signals sent through some kind of cable involving:

- electricity in copper wires, or
- infrared radiation in optic fibres.

We will first consider some of the issues involved with using (copper) wires. At its simplest, we may consider that data passes along a cable as a **signal** carried by a varying potential difference between two wires, with an associated electromagnetic wave between and around the wires. The two principal problems that arise are as follows.

- The p.d.s change as they travel longer distances via the cable. Because of the resistance and other electrical properties of the wire, the voltage signals get smaller and change shape. (There is more about *attenuation* and *dispersion* later in this chapter.)
- Electromagnetic waves spreading away from a cable can induce unwanted e.m.f.s in other cables (especially at high frequencies). Such unwanted signals are commonly called **electromagnetic noise** (interference) or 'crosstalk' if they come from other wires in the same cable.

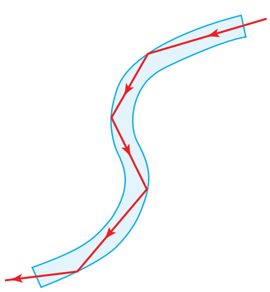
Electromagnetic noise can be greatly reduced by using **twisted pair cabling** in which the two insulated wires are twisted together (further explanation is not required). Figure 15.47 shows the kind of cable widely used in computer networking and in local telephone networks.



■ **Figure 15.47** Four-core twisted pair cabling

An alternative for reducing electromagnetic noise is the use of **co-axial cable**, which contains a central copper wire surrounded by an insulator and then an outer copper mesh – see Figure 15.48.

■ **Figure 15.48** Co-axial cable



■ **Figure 15.49** Rays following a curved path along an optic fibre

The mesh is connected to 0 V earth (grounded), meaning that electromagnetic signals cannot pass through it.

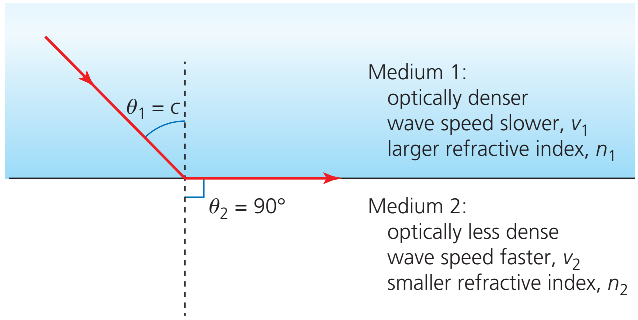
As we will explain, communication using optic fibres (which also transfer data using electromagnetic waves) can overcome the problems associated with using copper wires.

It was explained in Chapter 4 that total internal reflection can enable a light or an infrared beam to travel long distances along flexible glass fibres, as shown in the simplified representation in Figure 15.49. Because the fibres are thin (as thin as 0.01 mm), the angle of incidence of a ray is always larger than the critical angle, so that repeated internal reflections occur. Almost no light escapes from the fibre and, because it is possible to make glass of great clarity, the beam can travel very long distances without much absorption or scattering. The radiation used for communication is usually infrared because it is absorbed even less than visible light.

■ Total internal reflection and critical angle

This section summarizes ideas from Chapter 4. *Total internal reflection* can occur when waves meet a boundary with another medium in which they would have a faster speed – see Figure 15.50. When the angle of incidence, θ, equals the *critical angle*, c, the angle of refraction is 90° and the refracted ray is parallel to the boundary. If the angle of incidence is larger than the critical angle, *all* the radiation is reflected back into the original medium.

■ **Figure 15.50** The critical angle ($v_1 < v_2$; $n_1 > n_2$)



From the *Physics data booklet* (for Chapter 4) we know that:

$$\frac{n_1}{n_2} = \frac{v_2}{v_1} = \frac{\sin\theta_2}{\sin\theta_1}$$

At the critical angle, $\theta_1 = c$ and $\theta_2 = 90°$, so $\sin\theta_2 = 1$, and then:

$$\frac{n_1}{n_2} = \frac{1}{\sin c}$$

If the light is trying to pass from an optically denser material (medium 1) such as glass, plastic or water, *into air* (medium 2) then $n_2 = n_{air} = 1$, so that (replacing n_1 with n):

$$n = \frac{1}{\sin c}$$

This equation is given in the *Physics data booklet*.

Worked example

12 Infrared radiation is travelling along a glass optic fibre of refractive index 1.54 surrounded by air.

a What is the speed of the radiation?

b What is the smallest angle of incidence that the infrared rays can strike the boundary of the fibre at and still be totally internally reflected?

c How would your answer to (b) change if the fibre were surrounded by a different type of glass of refractive index 1.47 (instead of air)?

a $\frac{n_1}{n_2} = \frac{v_2}{v_1}$

$\frac{1.54}{1.0} = \frac{3.0 \times 10^8}{v_1}$

$v_1 = 1.9 \times 10^8\,\text{m s}^{-1}$

b $\frac{n_1}{n_2} = \frac{1}{\sin c}$, with $n_2 = 1$

$n_1 = \frac{1}{\sin c}$

$\sin c = \frac{1}{1.54} = 0.65$

$c = 40°$

c $\frac{n_1}{n_2} = \frac{1}{\sin c}$, with $n_2 = 1.47$

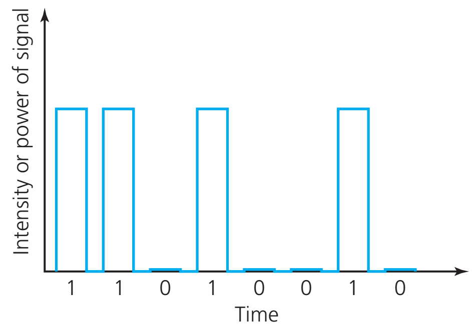
$\sin c = \frac{1.47}{1.54} = 0.95$

$c = 73°$

Nature of Science

Digital communication

Modern electronic communication is *digital* in nature; this means that rather than communicating using voltages or light/infrared intensities, which vary continuously (*analogue* signals), the data are transmitted as a very large number of pulses, each of which is intended to have only one of two possible levels (commonly called 0 and 1, or low and high). Figure 15.51



■ **Figure 15.51** Digital signal representing the binary number 11010010

represents a signal of only eight *binary* digits (*bits*), commonly called one *byte*. (The term **binary** describes a number in which each digit can only have one of two possible values.) This kind of digital signal can be sent along a cable as a series of voltage pulses, or pulses of light/infrared.

56 Discuss the factors that affect the choice between using electromagnetic waves in cables or electromagnetic waves in air for communicating data.

57 Suggest why electromagnetic noise (interference) is often a bigger problem at higher frequencies.

58 Explain, with the help of diagrams, why angles of incidence inside optic fibres will always be large if the fibres are thin. Assume that the original signal is transmitted parallel to the axis of the optic fibre.

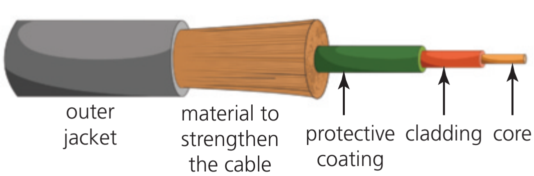
59 A typical optic fibre has a refractive index of 1.62.
- **a** What is the critical angle for such a fibre if it is surrounded by air?
- **b** What is the critical angle for such a fibre if it is surrounded by glass of refractive index 1.51?
- **c** Signals travel slower in glass of higher refractive index. Discuss whether or not this is a significant factor in choosing the type of glass used in an optic fibre.

60 Suggest why digital signals are used in preference to analogue signals for transferring data.

61 Suppose the digital signals shown in Figure 15.51 were transferred over a long distance and, as a result, the powers of the pulses were halved and the pulse times were doubled. Assuming that the pulses remain rectangular:
- **a** How would their energy have changed?
- **b** Sketch a graph of the received eight-bit signal.

■ Structure of optic fibres – cladding

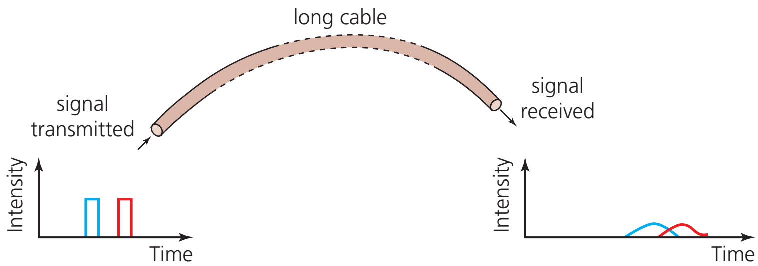
The sketch of an optic fibre shown in Figure 15.49 is much simplified compared with real optic fibres. It is very important that the surfaces of the fibres do not get damaged, or get any impurities on them, because the condition for total internal reflection would then change at such places, and some radiation could 'escape'. For this reason the fibres are surrounded by another layer of glass, known as *cladding* (Figure 15.52). The cladding protects the inner core fibre and also prevents the problems ('crosstalk') that would occur if different fibres came into contact with each other (multicore fibres are in common use). The cladding must have a refractive index lower than the inner fibre (see the Worked example 12 and question 59).



■ **Figure 15.52** A typical single core optic cable

■ Dispersion in optic fibres

When pulses (of various kinds) travel through long cables they each tend to spread out, so that they occupy longer and longer time intervals. This is known as *dispersion* and it is illustrated in Figure 15.53. (The two pulses also have reduced amplitudes, but we will discuss this later.) In this example the two transmitted pulses were clearly separate, but by the time they were received they overlapped to such an extent that the data they were transferring may not have been accessible. Of course, it is possible to increase the time between transmitted pulses in order to keep them separate, but that would reduce the amount of data that could be sent in any given time. Dispersion limits the rate at which data can be transferred.



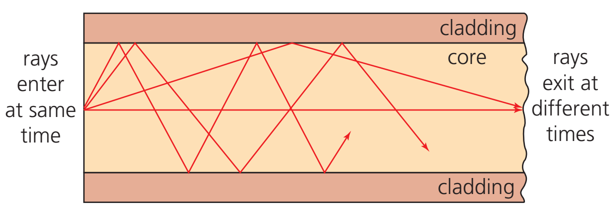
■ **Figure 15.53** Dispersion causes pulses to overlap

Waveguide and material dispersion in optic fibres

There are two principal causes of dispersion in optic fibres – *waveguide dispersion* and *material dispersion*.

Waveguide dispersion

If all the rays of light or infrared radiation that is transmitted along an optic fibre are parallel to begin with, they will follow parallel paths and take exactly the same time to reach any particular point along the cable. But this is an idealized situation and is not possible in practice. Rays representing a particular pulse can take slightly different paths and, therefore, different times to reach their destination. Figure 15.54 illustrates this problem (but it is exaggerated for clarity). This causes the spreading of pulses and the kind of dispersion known as **waveguide dispersion**, which is sometimes also known as 'modal dispersion', although this term will not be used in IB examinations. (An optic fibre is an example of a *waveguide*, which is a term used for any structure designed to transfer waves along a particular route.)

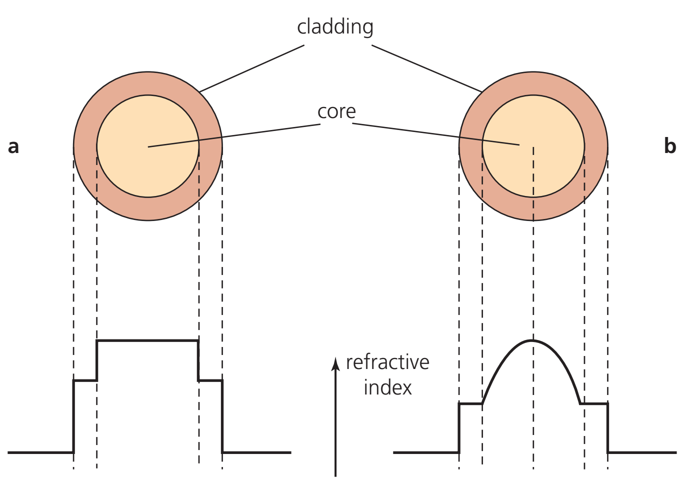


■ **Figure 15.54** Rays taking different paths and causing waveguide dispersion

To reduce the effects of waveguide dispersion it is better to use very thin fibres and to try to ensure that the light rays are parallel, and also to use *graded-index fibres*, as described next.

Step-index fibres and graded-index fibres

Up to this point in our description of optic fibres and their cladding we have assumed that they both have constant refractive indices, so that there is a sudden change ('step') of refractive index at the boundary between them. This simple arrangement is known as a *stepped-index fibre* and it is represented in Figure 15.55a.



■ **Figure 15.55** Cross-sectional variation of refractive indices in (a) step-index fibres and (b) graded-index fibres

Figure 15.55b represents a *graded-index fibre* in which the refractive index of the core optic fibre increases gradually towards a maximum at the centre. The effect of this is to relatively decrease the speed of the rays passing along the most direct (central) routes and relatively increase the speed of rays that pass closer to the outer surfaces of the core. The overall effect of having gradual changes in refractive index is to produce more central, curved paths, with less time differences between them, as shown in Figure 15.56. This reduces waveguide dispersion.

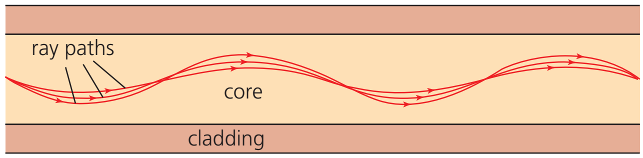


Figure 15.56 Typical paths of rays in a graded-index fibre

Material dispersion

Material dispersion is the name given to dispersion caused by the use of different wavelengths. In this respect it is a problem similar to chromatic aberration in lenses.

We know from Chapter 4 that the refractive index of a medium depends on the wavelength of the radiation. This effect is deliberately used in a prism to disperse light into a spectrum. For example, red light travels faster in glass (than the other visible colours), so that it has a slightly higher refractive index and it is the colour least deviated by passing through a prism. If different wavelengths (representing the same pulse) travel along the same path through an optic fibre, they will take slightly different times to reach their destination and will therefore produce dispersion.

The obvious solution to material dispersion is to use *monochromatic* radiation. Infrared LEDs are the most common source.

62 a Explain why infrared radiation, which is normally internally reflected, could pass between optic fibres if they came in contact with each other.
b Explain how the use of cladding will prevent this problem.

63 If the refractive indices of the cladding and the core in a stepped-index fibre are 1.60 and 1.55, respectively, what is the critical angle in the core?

64 Explain why dispersion limits the rate at which digital data can be communicated over longer distances.

65 Summarize, without reference to a diagram, how the use of graded-index fibres reduces waveguide dispersion.

66 Different data can be sent along the same optic fibre using different wavelengths of radiation. Discuss whether or not material dispersion affects this process.

Utilizations

Under-sea optic fibres

It is estimated that more than 95 per cent of all international communication and data are transferred using under-sea optical cables, which may already exceed one million kilometres in collective length. A typical cable (Figure 15.57) is about 7 cm in overall width and has a mass of about $10\,\text{kg}\,\text{m}^{-1}$.

Figure 15.57 An underwater optic fibre cable

These cables carry telephone conversations and are also the basic structure that enables the use of the internet worldwide. They are considered to be vital for the economic and social functioning of the modern world, and their importance is such that many countries consider them an important aspect of national security.

1 Search the internet to find an up-to-date map of the world's under-sea optic cables.

2 Discuss to what extent the length of under-sea cables might affect our speed of access to the internet.

Attenuation and the decibel (dB) scale

When anything (waves for example) is transmitted through a medium there will always be some *scattering* and *absorption*. **Scattering** occurs when waves are reflected by imperfections within the medium (usually randomly) and no longer continue to travel in their original direction. **Absorption** occurs when the wave energy is transferred to some other form within the material, usually internal energy. For these reasons the *intensity* of any signal reduces as it passes through a material.

This reduction of intensity, which is called *attenuation*, may or may not be significant under different circumstances. It should not be confused with the reduction in intensity always associated with waves that are *spreading out* from their sources.

Attenuation

Attenuation is the gradual loss of intensity of a signal as it passes through a material.

The high quality of the glass in optic fibres means that absorption and scattering should not be too significant over short distances, but they become important whenever long distances are involved. Scattering is the main cause of attenuation in optic fibres, but dispersion also affects attenuation. Consider again Figure 15.53. Even in the idealized example of no absorption or no scattering (so that the total power received is the same as transmitted, which is shown by equal areas) the intensity has still been reduced.

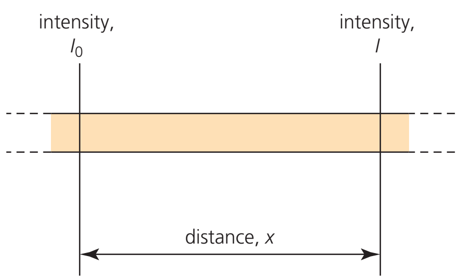


Figure 15.58 Variation of intensity with distance

It may be assumed that signal intensity is reduced by equal factors in equal distances, which means that intensity varies exponentially with distance. This means that we can quote a value for the attenuation per unit length of a system.

Calculating attenuation in decibels

The simplest way of calculating a value for attenuation would be to determine the ratio of the signal intensities (or powers) at two points, which would be I/I_0 as shown in Figure 15.58.

However, because of the exponential nature of the relationship, a logarithmic value is preferred:

$$\text{attenuation} = \log \frac{I}{I_0}$$

Attenuation calculated in this way is given the unit *bel*, but a smaller unit, the **decibel** (dB) is usually preferred:

$$\text{attenuation (dB)} = 10 \log \frac{I}{I_0}$$

This equation is given in the *Physics data booklet*. It can also be applied to powers:

$$\text{Attenuation (dB)} = 10 \log \frac{P}{P_0}$$

Because $I < I_0$, the attenuation in dB will be a negative number. We will not be using a symbol to represent attenuation in this topic.

Worked example

13 The attenuation in an optic fibre of length 100 km is −53 dB. If the input power is 0.0028 W, what power would be received:

a 100 km away?

b 200 km away?

a attenuation (dB) = $10 \log \frac{P}{P_0}$

$-53 = 10 \log \left(\frac{P}{0.0028}\right)$

$5.01 \times 10^{-6} = \frac{P}{0.0028}$

$P = 1.4 \times 10^{-8}$ W

b $5.01 \times 10^{-6} = \frac{P}{1.4 \times 10^{-8}}$

$P = 7.0 \times 10^{-14}$ W

It is common practice to quote attenuation per unit length, for example in dB km^{-1}. Table 15.1 gives some examples, but remember that this is just a rough guide because the values are also frequency dependent. (These numbers are sometimes called *attenuation coefficients*.)

Table 15.1 Typical attenuations per unit length

Type of cable	Attenuation (dB km^{-1})
twisted pairs (1 MHz)	50
co-axial (200 MHz)	100
optic fibre (10^{14} Hz)	1

Whatever kind of cable is used to communicate over long distances, attenuation will result in the signal intensity falling below an acceptable level unless the pulses can be amplified (and reshaped). The devices that perform this task are called *repeaters* or *regenerators*.

Summary of the advantages of optic fibres compared to wires

Optic fibres have many advantages over copper wires, especially whenever data needs to be transferred over long distances or at fast 'speeds'. This is because, when compared to copper wires of similar dimensions, optic fibre systems:

- have much lower attenuation (so that amplifiers/repeaters can be fewer and further apart)
- have much improved data transfer rates
- do not cause electromagnetic noise or crosstalk, nor are they affected by them from other cables
- are more secure (it is more difficult for third parties to access the data)
- are lighter in weight.

These advantages are so significant that optic fibres are the dominant means of transferring large amounts of data quickly over long distances. These are the 'superhighways' of communication. On a smaller scale, however, the convenience and lower overall cost of a copper wiring system may be more significant.

Nature of Science

The technology of optical communication

Total internal reflection is a relatively easily understood application of physics that was first described more than 400 years ago (by Kepler), although the concept of a beam of light being trapped inside a curved-shaped medium was not proposed until about 250 years later. The more recent application of this concept to the rapidly increasing use of optic fibres in the communication systems on which the modern world is so dependent is due to technological developments (such as improvements in the quality of glass), rather than new principles or new discoveries.

67 Sketch a graph to show how the intensity of a signal varies with distance as it travels along an optic fibre.

68 **a** What is the overall attenuation of a cable if the intensity of a signal is halved by passing through it?
b If a cable has an attenuation loss of 10 dB, by what percentage is the intensity of the output lower than the input?

69 The attenuation in a cable is rated at −0.36 dB for each 100 m. If the input power is 6.8 mW, what length of cable will reduce the power to 5.0×10^{-10} W?

70 The input power to a very long optic cable is 0.15 W. When the power falls below a certain value (*P*) the signal needs to be amplified/regenerated. Determine the value of *P* if the minimum distance between 'repeaters' is 80 km and the cable has an attenuation loss of 1.8 dB km^{-1}.

71 **a** Use the internet to determine the infrared frequency most commonly used in optical fibre communication.
b Why is this particular frequency chosen?

72 Suggest why attenuation in an optic fibre is frequency dependent.

73 The advantages of transferring data using glass rather than copper seem considerable. Use the internet to research possible reasons why copper wiring is still in widespread use.

15.4 (C4: Additional Higher) Medical imaging *– the body can be imaged using radiation generated from both outside and inside; imaging has enabled medical practitioners to improve diagnosis with fewer invasive procedures*

In recent years the rapid growth of computing power and technological advances have resulted in a dramatic increase in the number of utilizations of physics in medicine around the world. Nuclear medicine was mentioned in Chapter 12 and there are many applications of lasers, but in this section we will discuss the various physics principles that can be used to obtain images of bones, organs and tissues located inside the human body.

We begin with the use of X-rays and ultrasound, both of which involve sending penetrating waves into the body from outside, then we will discuss nuclear magnetic resonance.

Detection and recording of X-ray images in medical contexts

X-rays are useful in medical imaging because they are penetrating and some can pass deep into the body and out of the other side. However, some of the X-ray photons are absorbed, so when the transmitted X-rays are detected on the other side of the body, a 'shadow' or 'negative' picture can be produced. Figures 15.59a and 15.59b show how the bones have absorbed more of the X-rays than the rest of the hand. In this simple example, the X-rays are detected photographically (like light) by transferring the photons' energy to produce chemical changes in the film.

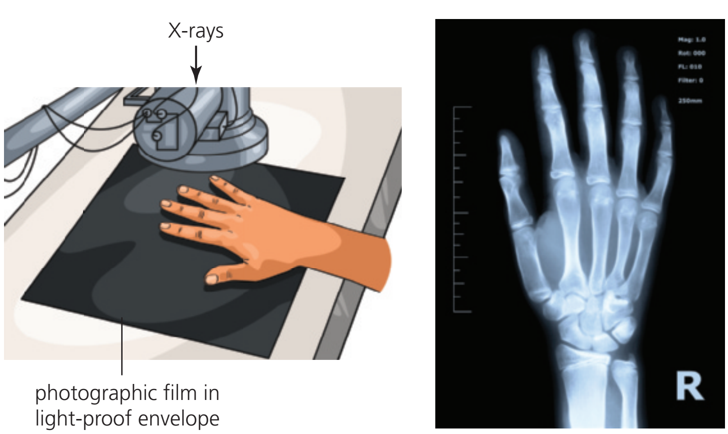


Figure 15.59 **(a)** Arrangement for X-raying a hand; **(b)** an X-ray of a hand

This technique is still widely used around the world. It is relatively inexpensive but the image must be chemically 'developed', which means there is a delay before the image is available to be viewed. Detection by **CCD (charge-coupled devices)** produces an immediate digital image and allows more control over the whole imaging and data-handling process. Equally as important, a detection process that requires a lower intensity will enable hospitals to use lower-power X-rays and reduce the health hazards associated with the use of X-rays.

Figure 15.60 shows the process of having a dental X-ray with the CCD in the patient's mouth.

■ **Figure 15.60** Having a dental X-ray taken

Nature of Science

Risk analysis

The benefits of using X-rays to diagnose illnesses are substantial and obvious. However X-rays are also a potential health risk because the energy carried by X-ray photons is high enough to cause ionization and possibly to cause damaging chemical and biological changes in the body (in a similar way to gamma ray photons, as discussed in Chapter 7).

The risk associated with directing a known amount of a particular kind of radiation into a particular patient cannot be known with certainty. Controlled experiments that involve exposing people (or animals) to radiation are clearly not acceptable, so the medical profession can only deal with statistics gathered indirectly from numerous previous events (medical or otherwise) in which people have been exposed to known, or estimated, amounts of radiation. Such data has been repeatedly analysed very extensively in order to assess the risk (the probability of harm) involved with any particular course of action.

Doctors must balance the risks of exposing a patient to X-rays against the medical benefits to be gained from a diagnosis or detailed knowledge of the medical problem that they need to treat. The health of medical staff involved with the use of X-rays in hospitals (*radiographers*) also needs to be considered. Standard safety measures include:

- using X-rays of as low a power as is consistent with their intended purpose
- using X-rays for as short a time as possible
- monitoring and limiting the number of X-rays taken of a patient
- preventing X-rays from going anywhere else other than the part of the body they are being used to examine.

The improvement in the technology involved with the production and, particularly, the detection of X-rays has been so considerable in recent years that the risks are now very well understood, well controlled and minimal. The required dose of radiation for any particular purpose is now so much reduced that a long trip in an aircraft (at high altitude) now involves a greater exposure to ionizing radiation than most X-rays.

■ Attenuation of X-rays

The amount of absorption depends on the energy carried by the X-ray photons and the type of material through which they are passing. We will now describe this in more detail, using the concept of *attenuation*, which we have already covered in our discussion of optic fibres.

X-rays of higher frequency have higher energy ($E = hf$) and are more penetrating. They are often described as having the *quality* of **hard X-rays** and they are produced in X-ray machines that use higher voltages. **Soft X-rays**, with lower photon energies, are more easily absorbed.

Attenuation is the gradual loss of intensity of radiation as it passes through a medium. In general, the principal reasons for attenuation are absorption and scattering. In the following discussion we will always assume that the incident radiation is all travelling in the same direction (a parallel beam).

For the energy of the X-ray photons commonly used in medical diagnosis, absorption due to the photoelectric effect is the principal means of attenuation and it is largely dependent on the proton number, Z, of the atoms present. For example, bone contains elements with a larger average proton number than soft tissue and therefore absorbs a higher percentage of X-rays.

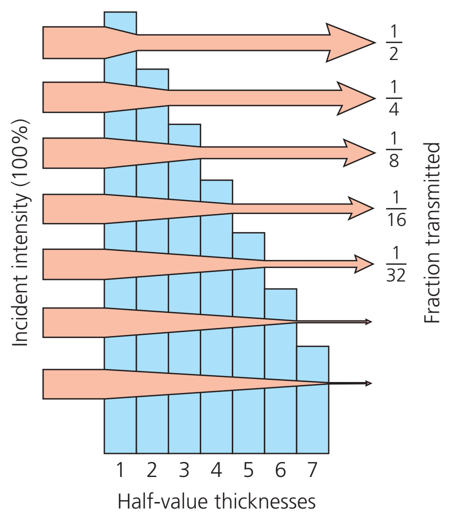
The attenuation can be expressed in decibels (as before, with optic fibres):

$$\text{attenuation (dB)} = 10\log\left(\frac{I}{I_0}\right)$$

This equation is given in the *Physics data booklet*.

We would normally expect that equal thicknesses of a homogeneous (uniform) medium would absorb (and scatter) the same percentage of the intensity, I. This is the essential characteristic of an exponential decrease and it can be simply characterized by referring to a **half-value thickness** (Figure 15.61). This is a similar concept to half-life in radioactive decay.

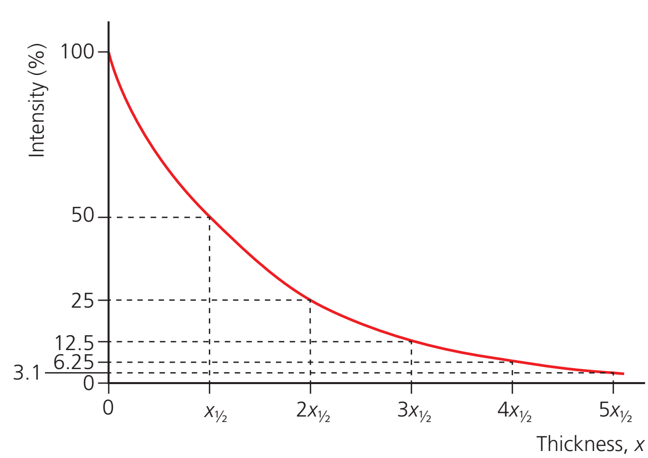
Figure 15.61 Half-value thicknesses



Half-value thickness, $x_{1/2}$, is defined as the thickness of a medium that will reduce the transmitted intensity to half its previous value.

Figure 15.62 shows how this can be represented graphically.

Figure 15.62 Graph of intensity–thickness



X-ray photons with higher energy will be more penetrating, so their half-value thickness of a particular medium will also be larger.

Worked example

14 **a** When directed through a homogeneous material, the intensity of an X-ray beam falls to $\frac{1}{8}$ of its initial value over a distance of 15 cm. What is its half-value thickness?

b What overall thickness would be needed to reduce the intensity to $\frac{1}{16}$?

c How would the half-value thickness of this material change if X-rays of longer wavelength were used?

a 1 to $\frac{1}{2}$ to $\frac{1}{4}$ to $\frac{1}{8}$ requires three half-thicknesses, so one half-thickness = $\frac{15}{3}$ = 5 cm

b $\frac{1}{8}$ to $\frac{1}{16}$ requires one more half-thickness: 15 + 5 = 20 cm

c Photons of a longer wavelength have less energy, so they would be less penetrating and the half-thickness would be smaller.

An exponential equation for attenuation

An exponential decrease like that shown in Figure 15.62 can be represented by an equation of the form:

$$I = I_o e^{-\mu x}$$

where I_o and I are the intensities before and after the attenuation caused by passing through a medium of thickness x. (We have used similar equations in the study of radioactive decay and capacitor discharge.) This equation is given in the *Physics data booklet*.

μ is a constant called the *linear* **attenuation coefficient**. It represents the amount of attenuation in a particular medium (for radiation of a specified wavelength).

A larger value of μ represents more attenuation and corresponds to a smaller value for the half-value thickness. Both of these properties vary significantly with the energy of the X-ray photons.

Taking logarithms and rearranging, we get:

$$\mu = \frac{1}{x} \ln\left(\frac{I_0}{I}\right)$$

The unit of μ that is normally used with X-rays is cm^{-1}. A value for the attenuation coefficient of a specific material can be found by using values of I_o and I for a known thickness, x. Preferably a range of values would be used and the attenuation coefficient determined from a suitable graph.

The linear attenuation coefficient and half-value thickness are different ways of expressing the same information. Their relationship can be derived by putting $I = I_o/2$ for $x = x_{1/2}$ in the above equation:

$$\mu = \frac{1}{x_{1/2}} \times \ln\left(\frac{2I_0}{I_0}\right)$$

or:

$$\mu x_{1/2} = \ln 2$$

This equation is given in the *Physics data booklet*.

Attenuation coefficients are often called absorption coefficients although, as we have seen, absorption is not the only cause of attenuation.

Worked examples

15 An X-ray beam is reduced to 93 per cent of its intensity when it passes through a medium of thickness 0.48 mm.

a What is the linear attenuation coefficient of the medium?
b What is its half-value thickness?

a $\mu = \frac{1}{x}\ln\left(\frac{I_0}{I}\right)$

$= \left(\frac{1}{0.48}\right) \times \ln\left(\frac{I_0}{0.93I_o}\right)$

$= 0.15\,\text{mm}^{-1}$

b $\mu x_{1/2} = \ln 2$

$x_{1/2} = \frac{0.693}{0.151} = 4.6\,\text{mm}$

16 An X-ray beam of intensity 580 mW m^{-2} is passed through parallel layers of two different materials. The first layer is 4.3 cm thick and has a linear attenuation coefficient of 0.89 cm^{-1}; the second layer is 1.3 cm thick with a linear attenuation coefficient of 0.27 cm^{-1}.

a Determine the intensity of the X-rays that cross the boundary between the layers.
b What intensity emerges from the second layer?

a $I = I_o e^{-\mu x}$

$= 580e^{-0.89 \times 4.3}$

$\ln I = \ln 580 - (0.89 \times 4.3)$

$I = 13\,\text{mWm}^{-2}$

b $I = I_o e^{-\mu x}$

$= 13e^{-0.27 \times 1.3}$

$\ln I = \ln 13 - (0.27 \times 1.3)$

$I = 9.1\,\text{mWm}^{-2}$

Mass attenuation coefficient

The *linear* attenuation coefficient of a medium represents how the intensity of an X-ray beam decreases with unit distance travelled through the medium, but the *mass* attenuation coefficient is often more useful when comparing different materials. The mass attenuation coefficient represents how the intensity of a beam decreases as it passes through unit mass of a particular medium.

$$\text{mass attenuation coefficient} = \frac{\text{linear attenuation coefficient}}{\text{density}} = \frac{\mu}{\rho}$$

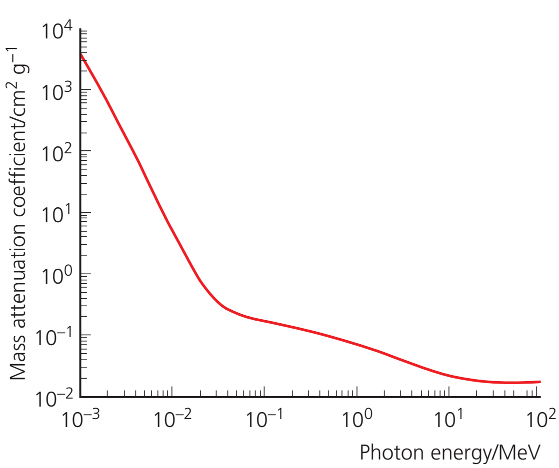


Figure 15.63 Mass attenuation coefficient of water: variation with X-ray photon energy

This equation is *not* given in the *Physics data booklet*. A mass attenuation coefficient most commonly has the unit cm^2 g^{-1}.

Figure 15.63 show how the mass attenuation coefficient of water varies with photon energy. You should be able to show that a typical X-ray wavelength of 1×10^{-10} m has a photon energy of 1.2×10^{-2} MeV, which corresponds approximately to a mass attenuation coefficient of 0.05 cm^2 g^{-1}.

Worked example

17 Use Figure 15.63 to estimate the *linear* attenuation coefficient of water for photons of energy 100 keV.

For 0.1 MeV photons, the mass absorption coefficient is approximately $2 \times 10^{-1}\,cm^2\,g^{-1}$.

Mass attenuation coefficient = $2 \times 10^{-1} = \frac{\mu}{\rho}$, and the density of water is $1.0\,g\,cm^{-3}$.

$\mu = 2 \times 10^{-1} \times 1.0 = 0.2\,cm^{-1}$

Example of X-ray attenuation in the human body

Figure 15.64 shows a simplified cross-section of a human thigh and the relative intensities of X-rays passing through it. It has a bone (femur), which has a linear absorption coefficient of $0.62\,cm^{-1}$, surrounded by soft tissue with an absorption coefficient of $0.12\,cm^{-1}$. X-rays following path B emerge from the thigh with their intensity reduced to about 15 per cent; X-rays that follow path A, passing through the bone, have their intensity reduced to about 1 per cent. In this example, the overall attenuation of the X-rays passing through the bone (and soft tissue) is about 15 times higher than for the X-rays that only pass through the soft tissue.

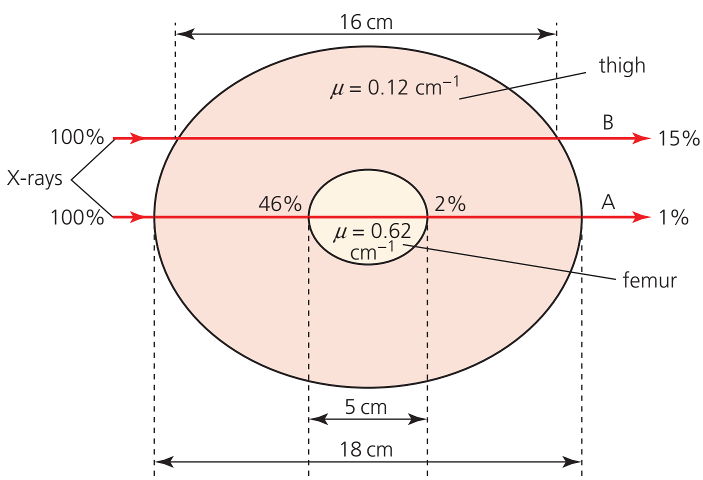


Figure 15.64 Relative intensities of X-rays passing through a thigh

74 A parallel X-ray beam was reduced in intensity from $100\,W\,m^{-2}$ to $74\,W\,m^{-2}$ after passing through 3.0 cm of a medium. Calculate the linear attenuation coefficient in cm^{-1}.

75 A parallel beam of X-rays passes though a material with a half-value thickness of 3.7 cm. If the incident beam has an intensity of $150\,W\,m^{-2}$, calculate:
- **a** the linear attenuation coefficient
- **b** the percentage of the incident intensity that emerges from a 4.5 cm thickness of the material.

76 The intensity of a parallel X-ray beam is reduced from $195\,W\,m^{-2}$ to $127\,W\,m^{-2}$ when it passes through a medium of thickness 2.10 mm.
- **a** What is the total power of the incident beam if it covers an area of 16.0 cm by 20.5 cm?
- **b** If the wavelength is 2.27×10^{-11} m, how many photons enter the medium every second?
- **c** Calculate the linear attenuation coefficient.
- **d** What is the half-value thickness of the medium?
- **e** If the accelerating voltage producing the X-rays is increased, suggest how the half-value thickness will change.

77 A medium has a density of $1.9\,g\,cm^{-3}$ and a linear attenuation coefficient of $0.22\,mm^{-1}$.
- **a** What thickness of this medium will reduce a parallel X-ray beam to 33 per cent of its incident intensity?
- **b** What is the mass attenuation coefficient of the medium?

78 Explain why it is reasonable to expect that increasing the thicknesses of a medium by equal amounts will result in equal percentage falls in the transmitted intensities of X-rays.

79 A material has a density of $1.3\,g\,cm^{-3}$ and a mass attenuation coefficient of $2.1\,cm^2\,g^{-1}$. Calculate its half-value thickness.

80 If the linear attenuation coefficient for certain X-rays in soft tissue is quoted at $0.35\,mm^{-1}$, what is the linear attenuation coefficient for bone (with the same X-rays) if bone has a half-value thickness that is 150 times larger than for soft tissue?

81 Confirm the values quoted in Figure 15.64 for the relative intensities of the X-rays passing through the thigh.

82 Data showing the relative intensity for X-rays passing through a certain material is shown in Table 15.2.

Thickness (cm)	Relative intensity
0.0	1.00
1.0	0.78
2.0	0.58
3.0	0.48
4.0	0.33
5.0	0.27
6.0	0.20
7.0	0.17

Table 15.2 Data on X-ray absorption

a Draw a relative intensity–thickness graph to represent this attenuation.
b Draw a graph of the logarithm of the relative intensity against thickness.
c Use your two graphs to determine the attenuation coefficient and the half-value thickness.

83 Why is it reasonable to assume that both X-rays and gamma rays of the same energy will have the same value of half-value thickness?

Obtaining good quality images with X-rays

X-rays are not *focused* to form an image, so it is important for **sharp images** that all the radiation reaching any particular place on the film has followed the same path from the X-ray source and through the patient. When we describe an image as *sharp* we mean that any 'edges' between different parts of the image are distinct and precise, and that it has good resolution. The source of X-rays needs to act like a point source, as in Figure 15.65a, rather than an extended source, as shown in Figure 15.65b.

In 15.65a X-rays can only have followed one particular path to reach point P, but in 15.65b X-rays can arrive at point P from different directions. In the same way, a point source of light is needed in order to form sharp shadows.

For *sharp* images, ideally the film or detector should also be as close as possible to the patient, who should not move during the exposure time, and the X-ray source placed as far away as possible. Of course, increasing the distance between the patient and the source has the significant disadvantage of requiring a source of higher intensity, and/or a longer exposure time.

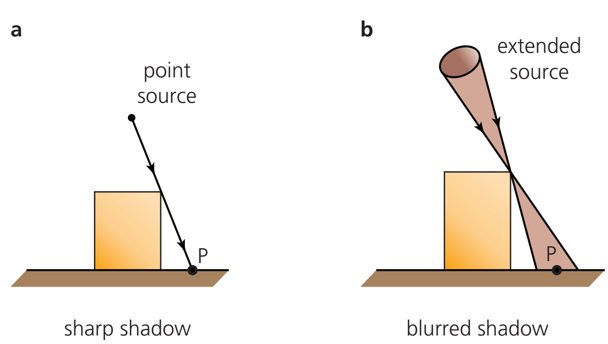


Figure 15.65 Point and extended sources

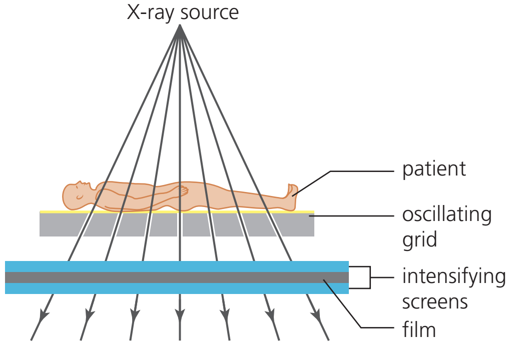


Figure 15.66 X-ray arrangement with the film placed between two intensifying screens

To increase the brightness of an image formed on photographic film, an **intensifying screen(s)** can be used to enhance the image. These screens contain **fluorescent** materials that emit light when struck by X-rays. Photographic film is much more sensitive to light than X-rays, so the image is intensified. Figure 15.66 shows a possible arrangement, with the film placed between two intensifying screens.

If X-rays *scattered* from all parts of the body can reach the film, the **contrast** of the image will be reduced because all parts of the film will receive more X-rays than they would if only X-rays travelling directly from the source reached the film. This effect can be reduced by using a *collimating* grid as shown in Figure 15.66. A **collimator** makes all rays passing through it parallel to each other. The grid has to be oscillated from side to side during the exposure to enable all relevant parts of the patient to be imaged.

For some applications, the contrast of an X-ray image can also be enhanced by temporarily introducing into the patient's body something that will affect the absorption of X-rays. For the digestive system this could be a (non-poisonous!) substance that the patient has to drink. For CT scans (see below) a 'contrast medium' is sometimes injected into the bloodstream.

The sharpness of an image can also be digitally enhanced later by a computer program.

Computed tomography (CT)

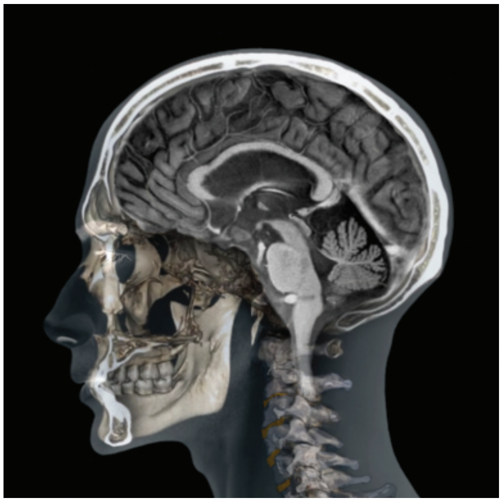
This technique uses computer control to obtain detailed, sharp, layered images of high contrast.

The X-ray images described so far have all been two-dimensional. Detailed images of three-dimensional objects can be obtained with more sophisticated and expensive equipment that is controlled by computers. **Tomography** is the term used to describe obtaining images of a three-dimensional object as a series of sections or 'slices'.

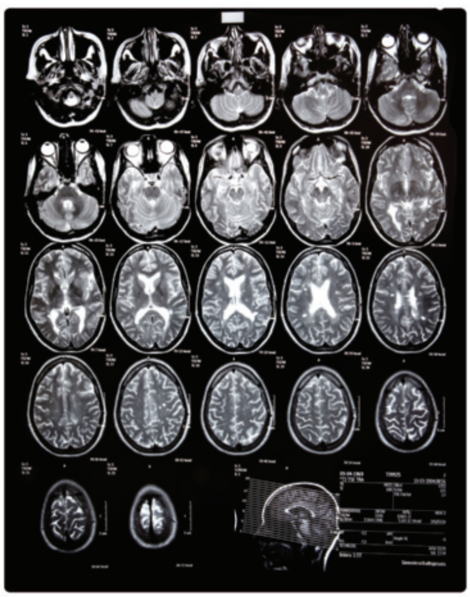
The principles of tomography are not difficult to understand, but **CT scans** (also called *CAT scans*, computed axial tomography) of anything as complicated as a human body (or part of it) require considerable computing power and expensive equipment. Figure 15.67 shows a patient in a CT scanner.

For some people a CT scan is an uncomfortable experience – the body must remain as stationary as possible while the source of X-rays and an array of detectors are rotated around in

Figure 15.67 Patient in a CT scanner



■ **Figure 15.68** CT scan of head



■ **Figure 15.69** CT scans of 'slices' within a head

a precise plane. What any particular detector receives will vary as the angle of incidence on the body changes. An image is not obtained directly, but the power of the computer is used to analyse all the data received during the rotation and to construct an image of the body in that particular plane. Other planes are then scanned by slightly moving the bed on which the patient is lying. It is then possible to use the computer to obtain images of any plane, or to obtain a complete three-dimensional image. Like digital images obtained with a camera, the images may be enhanced and changed in a wide variety of ways.

Using the vast quantity of data collected, CT scans provide a range of high-contrast images with good resolution. They are able to distinguish tissues with a density difference as low as 1 per cent. Undoubtedly they provide images more useful than simple X-rays, but CT scanners are expensive to buy and to operate, and the longer exposure time compared with X-rays increases the radiation dose received by a patient by up to 1000 times, which is a significant additional health risk that has to be considered by the doctor. There has been a rapid increase in worldwide use of CT scanners in recent years.

Figure 15.68 shows a composite image, which should be compared with Figure 15.69, which shows CT scans of individual 'slices'.

84 Make a list of the advantages of using electronic detectors rather than detecting X-rays photographically.

85 Explain why the detection of scattered X-rays reduces the contrast of an X-ray image.

86 An X-ray imaging system was redesigned to increase the distance between the source and the patient.
  a Suggest a reason why this was done.
  b If the same source is used, why will the intensity reaching the patient be reduced?

87 Find out what is meant by a 'barium meal'.

88 List two advantage and two disadvantages of CT scans compared to conventional X-rays.

Ultrasound imaging

High-frequency sound waves can be used as an alternative to electromagnetic waves (X-rays) for obtaining images from inside the body and diagnosing illnesses. Any sound wave that has a frequency higher than that which can be heard by humans (20 kHz) is called **ultrasound** and is described as *ultrasonic*. Frequencies used in ultrasonic imaging are typically between 2 MHz and 20 MHz.

Figure 15.70 Abdominal ultrasound scan (B-scan)

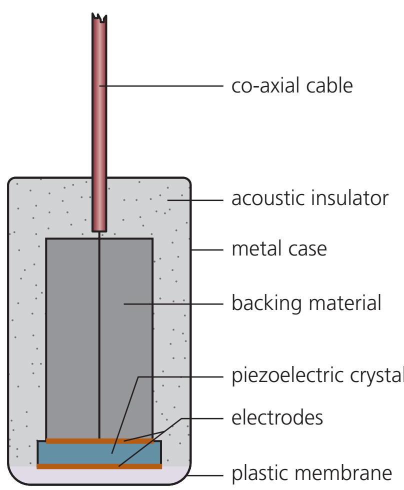
Unlike X-rays, ultrasound is not an ionizing radiation and has no known health risk. The equipment is also generally cheaper and easier to use, however the images produced do not usually have the same detail (high resolution) as those produced by CT scans because the longer wavelengths of ultrasound reduce resolution because of greater diffraction. Ultrasound techniques can produce better images of some soft tissues than X-rays, however.

The ultrasound waves are usually directed into the patient using a handheld **transducer** (often called a **probe**) which converts electrical signals into ultrasound waves. Reflections are received back at the same device (Figure 15.70). In general, some waves will be reflected whenever they arrive at any boundary between two different media. This will be discussed in more detail later, but first we will look at how the ultrasonic waves are produced and detected.

Generation and detection of ultrasound in medical contexts

When certain materials are under stress (stretched or squashed) a potential difference is induced across them. Conversely, when the same material has a p.d. applied to it, a (very small) change of shape is produced. This is called the **piezoelectric effect**. Quartz crystals are commonly used in piezoelectric transducers and are ideal for the production and detection of ultrasound. An *alternating* p.d. applied across a piezoelectric crystal will make its surfaces oscillate at the same frequency; this disturbs the surrounding medium and a longitudinal ultrasonic wave propagates away. See Figure 15.71.

Figure 15.71 Piezoelectric transducer



When reflected ultrasound waves are incident on the same crystal, they can be detected by the alternating p.d. induced, which has the same frequency as the waves.

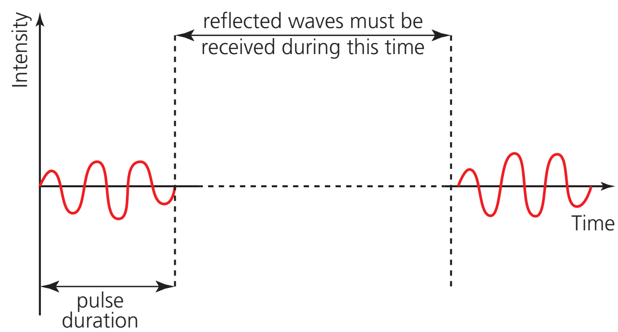
Basic principles of ultrasound imaging

Ultrasonic waves are directed into the patient's body and reflections travel back to the transducer whenever the waves meet a boundary between two different media. If the original direction of the waves is known, as well as the speed of the waves and the time delay between the emitted and reflected waves, then the location of where the reflection occurred can be pinpointed. (This is very similar to the principles used in the *echolocation* systems of *sonar* and *radar*.)

The ultrasound waves cannot be emitted and reflected continuously, because then there would be no easy way of knowing which waves caused which reflections. For this reason, the ultrasound is transmitted in short *pulses*, and the time between them should be longer than the longest time it could take a reflection to be received back at the transducer. Typically the time between pulses is about 1×10^{-4} s, which means that the *pulses* of ultrasound have a frequency of about 10 kHz (**pulse repetition frequency**). Remember that the ultrasound waves themselves have a much higher frequency of about 10^4 kHz.

Each pulse might typically contains two waves or more, so that a typical *pulse duration* is $2 \times 1/f = 2 \times 1/(1 \times 10^7) = 2 \times 10^{-7}$ s – see Figure 15.72. This means that the time between pulses is about 500 times longer than the duration of each pulse. Longer pulse durations improve the resolution of images.

Figure 15.72 Pulses of ultrasound



Worked examples

18 Ultrasound waves travelling at an average speed of $1600\,m\,s^{-1}$ through a person's body reflect off an organ and are received back at the probe after 32 μs.

a What is the distance of the surface of the organ beneath the skin?

b What assumption was made in this calculation?

a distance × 2 = speed × time = $1600 \times 32 \times 10^{-6}$

distance = 2.6×10^{-2} m (2.6 cm)

b The waves travel perpendicularly to the surface.

19 To examine structures relatively far from the surface of the body, the pulse repetition frequency may need to be reduced to, for example, 2 kHz in order to increase the time between pulses.

a What total distance can an ultrasound wave travelling at an average speed of $1580\,m\,s^{-1}$ travel in the time between pulses (assume that the pulse duration is negligible)?

b Estimate the number of pulses that could be in the body at any one instant.

a distance = speed × time = 1580 × (1/2000) = 0.79 m (79 cm)

b The maximum distance a (reflected) wave could travel is approximately twice the width of the body, which might be about 70 cm (depending on orientation). Under those circumstance there would only be one pulse in the body at any time. That is, the reflected pulse would be received before the next pulse was emitted.

Acoustic impedance

In general terms, **acoustic impedance** is a measure of the opposition of a medium to the flow of sound through it. Knowledge of acoustic impedance is needed to understand ultrasound imaging because the reflection of ultrasound waves from boundaries between media depends on how their acoustic impedances compare.

The bigger the difference in impedances, the higher the percentage of incident waves that are reflected.

Acoustic impedance, Z, can be defined as:

acoustic impedance = density of substance × speed of sound in that substance

Or, in symbols:

$$Z = \rho c$$

This equation is given in the *Physics data booklet*.

The units of impedance are $kg\,m^{-2}\,s^{-1}$. Table 15.3 provides a list of some acoustic impedances relevant to ultrasound imaging at a typical frequency. (Acoustic impedance is a frequency-dependent property.)

Table 15.3 Acoustic properties of parts of the human body

Medium	Speed, v/m s^{-1}	Density, ρ/kg m^{-3}	Acoustic impedance, Z/10^6 kg m^{-2} s^{-1}
air	340	1.2	0.000408
fat	1460	950	1.39
water	1480	1000	1.48
soft tissue	1500	1050	1.58
kidney	1040	1560	1.62
liver	1060	1570	1.66
blood	1575	1057	1.66
muscle	1580	1080	1.71
skin	1730	1150	1.99
bone	4080		7.79

Ultrasound imaging clearly depends on the reflection of the waves from various boundaries, but one place where reflection is definitely *not* required is at the point where the waves enter the patient's body. From Table 15.3 we can see clearly that the acoustic impedance of air is very much lower than that of skin, which means that the percentage reflected from the skin would be very high if there was an air gap. Therefore, the transducer must be in good contact with the skin and this is helped by the use of a gel (coupling medium) between them. The gel has an acoustic impedance similar to that of skin. This is an example of a process known as *impedance matching*.

The acoustic impedance of steel is about $45 \times 10^6\,kg\,m^{-2}\,s^{-1}$. When this is compared with air ($0.000\,408 \times 10^6\,kg\,m^{-2}\,s^{-1}$), it is clear why sound waves in air reflect off steel well.

Worked example

20 a The speed of ultrasound waves in a particular part of a patient's body is $1580\,m\,s^{-1}$. If the tissue has a density of $1050\,kg\,m^{-3}$, what is its acoustic impedance?

b Use Table 15.3 to determine an average density for bone.

c Apart from air, which pair of media in the table have the highest percentage reflection?

a $Z = \rho c = 1050 \times 1580 = 1.66 \times 10^6\,kg\,m^{-2}\,s^{-1}$

b $Z = \rho c$

$7.79 \times 10^6 = \rho \times 4080$

$\rho = 1910\,kg\,m^{-3}$

c Bone and fat, because their acoustic impedances have the greatest difference.

89 The acoustic impedance of a certain material is $2.08 \times 10^6\,kg\,m^{-2}\,s^{-1}$. What is the speed of sound waves in this material if its density is $1250\,kg\,m^{-3}$?

90 A pulse contains three complete waves that have a frequency of 2.0 MHz. They travel at a speed of $1510\,m\,s^{-1}$.
- **a** What is the duration of the pulse?
- **b** If the pulse repetition frequency is 8.4 kHz, how far can the waves in a pulse travel before the next pulse is emitted?
- **c** Are these frequencies suitable for the examination of a part of the body that is 10 cm below the surface of the skin?

91 The ratio of reflected intensity, I_r, to incident energy, I_o, at a boundary between two media of acoustic impedances Z_1 and Z_2 is given by the following equation (which is *not* required in this course):

$$\frac{I_r}{I_o} = \frac{(Z_1 - Z_2)^2}{(Z_1 + Z_2)^2}$$

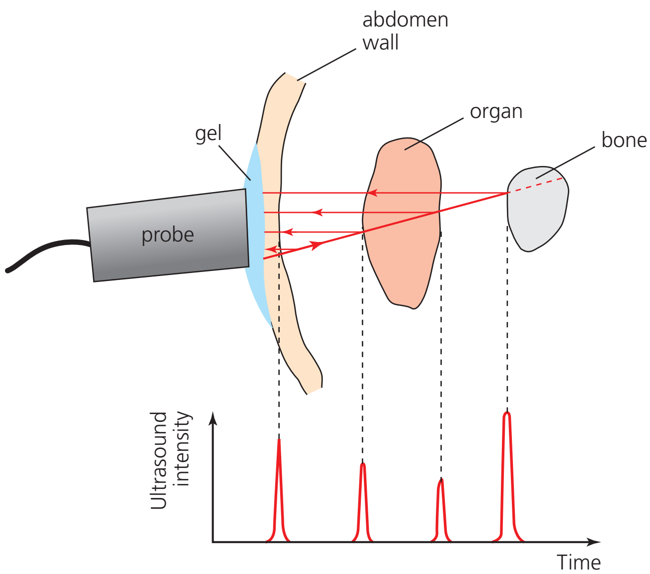
- **a** Show that this equation predicts that all of the incident energy is transmitted if the two media have equal impedances.
- **b** What percentage of the intensity is reflected if one medium has twice the impedance of the other? Does your answer depend on the direction in which the waves are travelling?
- **c** If ultrasound of intensity $0.1000\,W\,cm^{-2}$ is incident on the boundary between soft tissue and liver, what intensity is transmitted into the liver?
- **d** Estimate the percentage of intensity transmitted into skin when ultrasonic waves are incident on it from air. Hence, explain why gels are used with ultrasound transducers.

92 Why would you expect X-ray imaging to produce better resolution than ultrasonic imaging?

A-scans and B-scans

The simplest type of ultrasound scan is an **A-scan**. At each boundary between different media (for example, fat–muscle or muscle–bone) some of the waves are reflected and some are transmitted. The reflected waves are received by the piezoelectric transducer and a p.d. is induced that can be displayed as a p.d.–time graph. Figure 15.73 shows a typical example. The amplitudes (intensities) of the reflected pulses received at the transducer depend on the distance that the waves have travelled, the type of boundary from which they have been reflected and the number of other media boundaries that the waves have crossed. It is called an A-scan because it displays information in the form of varying *amplitudes*.

A-scans are useful for obtaining accurate measurements of a known situation. By moving the transducer to different locations the exact position, size and shape of an organ can be determined. There are a wide range of applications of this technology outside of medicine – for example in the detection of faults in pipes and railway lines.



■ **Figure 15.73** Reflected waves in an A-scan (This time scale is not regular)

Worked example

21 Consider Figure 15.73. Calculate the width, *s*, of the organ if the time delay between reflections from the two boundaries is 73 μs.

$2s = v\Delta t = 1040 \times 73 \times 10^{-6}$ (assume a wave speed of $1040\,m\,s^{-1}$)

$s = 0.038\,m$ (3.8 cm)

B-scans are more widely used than A-scans in hospitals. These display the information in the form of varying brightness in a two-dimensional, real-time video image such as that shown in Figure 15.70 (A-scans are one-dimensional). They are called B-scans because they display information in terms of *brightness*.

The information is obtained in essentially the same way as in an A-scan, except that the amplitude of the reflected wave is represented by the brightness of a dot on a screen. The picture is constructed by a computer program using the information from one or more transducers inside the ultrasound probe, transmitting waves in a slightly different direction while the probe is moved to different positions.

Ultrasound scan frequency

The range of frequencies used in ultrasound imaging has already been mentioned (between 2 MHz and 20 MHz), but *why* are those frequencies used? The choice is affected by two wave properties, which are frequency dependent in opposite senses – diffraction and attenuation.

1 It is important that the ultrasound beam emerging from the transducer is parallel and can be directed towards a particular location on the patient's body. This means that there should not be much *diffraction* of the emerging beam. Therefore, the wavelength needs to be significantly less than the aperture on the transducer from which it is transmitted, because the amount of diffraction depends on the ratio λ/b, where b is the width of the structure causing the diffraction. For similar reasons, for good *resolution* the wavelength also needs to be much smaller than the parts of the body being examined (Chapter 9).
So, in order to reduce the possible effects of diffraction and to improve resolution, a high frequency (short wavelength) is preferable.
Because the dimensions of the transducer and parts of the body may be typically a few millimetres or more, the chosen wavelength needs to be shorter than approximately 1 mm, which corresponds to a minimum frequency of about 2 MHz (using $v = 1600\,\text{m s}^{-1}$), although higher frequencies will be preferable in this respect.

2 It is also important that as little of the ultrasound energy is absorbed as possible. Because attenuation increases with frequency, *lower* frequencies are preferable in this respect. See Figure 15.74, which is a simplification but broadly represents the situation.

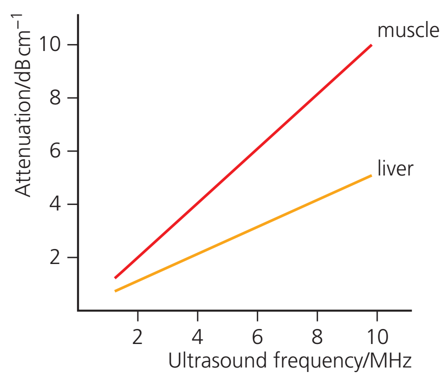


Figure 15.74 How attenuation of a parallel ultrasound beam varies with frequency in two parts of the body (simplified)

For any particular examination, the ultrasound frequency used depends on the acoustic impedances of the part(s) of the body being scanned. The duration of the pulse is another factor affecting the resolution of the image (longer pulses are preferred). The pulse repetition frequency may also need to be adjusted for parts of the body that are at different distances from the surface of the skin. The ultrasound technician, or the doctor, may need to make adjustments depending on the particular circumstances.

Summary of advantages and disadvantages of using ultrasound for medical diagnosis

Advantages:

- No known harmful effects on the body.
- Relatively inexpensive.
- Equipment is mobile and can be moved easily to different locations.
- Non-invasive – does not involve opening the body or inserting anything into the body.
- Particularly useful for examining the boundaries between soft tissues, where there may be only slight differences in density.
- Images are available in real-time and may also be seen at that time by the patient.

Disadvantages:

- Poor resolution because of the relatively long wavelengths used.
- Ultrasonic waves do not transmit well through bone.
- Cannot be used effectively with spaces that contain gas (such as the lungs and stomach) because the waves are strongly reflected at the boundaries.

Utilizations

Ultrasound and ophthalmology

Ultrasound, both A- and B-scans, is particularly useful for the examination of the soft tissues and fluids of the eye. (*Ophthalmology* is the branch of medicine that deals with the eye.)

One widely used application of the A-scan is a quick and accurate measurement of the dimensions of the eyeball. The ultrasonic waves usually have a frequency in the range of 8–12 MHz. Information about the dimensions of the eye may be useful in diagnosing ocular problems and is very helpful for calculating the power needed for an artificial lens implant, such as in the treatment of cataracts.

B-scans are used to obtain a visualization of the internal structure of the eye (see Figure 15.76), particularly the retina at the back of the eye. For these scans the waves can be directed into the eye through closed eyelids, as shown in Figure 15.75.

Figure 15.75 Ultrasonic examination of the eye

Figure 15.76 Ultrasound image of an eye

1 Sketch the structure of the eye seen in Figure 15.76 and label the different parts (research any information you may need).

ToK Link

We often only see what we expect to see

'It's not what you look at that matters, it's what you see.' – Henry David Thoreau. To what extent do you agree with this comment on the impact of factors such as expectation on perception?

An untrained observer looking at images that represent the inside of the human body often find it difficult to interpret the true meaning of the picture they are looking at. Doctors need to be carefully trained in these skills. But we all tend to see in any image what we *expect* to see because the brain does not have the time to fully evaluate all the information that is available, and it may jump to (sometimes incorrect) conclusions based on previous observations. Figure 15.77 provides a simple example. Of course, when we are alerted to the fact that an image requires special scrutiny, we give it much more attention than it would otherwise receive.

Figure 15.77 There are 12 faces in this picture – can you spot them?

93 Consider Figure 15.73.
- **a** Explain how it is possible for the reflected waves that have travelled the longest distance to have the largest amplitude.
- **b** Explain why the time scale is 'not regular'.
- **c** If the width of the organ was 2.2 cm and the wave speed in it was 1030 m s^{-1}, what was the time difference between the second and third reflected pulses?
- **d** Determine the average acoustic impedance of the organ if its density was 1540 kg m^{-3}.

94 **a** Explain why the use of a gel with an ultrasound probe can be considered as an example of impedance matching.
- **b** Suggest a suitable value for the acoustic impedance of the gel.

95 Consider Figure 15.74. Estimate the percentage of the incident ultrasound intensity that is transmitted after passing through 2 cm of liver if the frequency is:
- **a** 4 MHz
- **b** 8 MHz.

96 **a** Suggest a reason why ultrasound may be of little use in the diagnosis of problems with the brain.
- **b** Give two reasons why ultrasound is used in preference to X-rays for pre-natal scanning.

97 Find another widespread use of ultrasound for medical diagnosis. Why is ultrasound used for this (rather than other imaging techniques)?

98 Suggest why ultrasound is not used for imaging lungs.

99 Research the use of Doppler ultrasound for diagnosing some heart conditions.

Nuclear magnetic resonance

Resonance is the name given to the effect in which a system (that can oscillate) absorbs energy from another external oscillating source. Resonance effects are greatest when the external source has a frequency equal to the natural frequency of the system. Pushing someone on a swing is an easily understood example – the system (the swing) gains energy (its amplitude increases) if it is pushed at the same frequency as it swings naturally on its own.

Nuclear magnetic resonance (NMR) provides an alternative to CT scans for providing images of sections through the body. Medical applications of NMR are also commonly called **magnetic resonance imaging** (MRI, avoiding the use of the emotive term 'nuclear').

NMR is particularly useful for brain scans because it is better than CT scans at displaying images of soft tissues. Its major advantage is that no dangerous radiations are used, but NMR scanners are more expensive than using X-rays and take a longer time to produce an image. Instead of sending penetrating and potentially dangerous X-rays into the body, NMR involves getting protons in the human body to absorb and then re-emit energy from an oscillating electromagnetic field. This is a complicated process and the following is only an outline of what is involved.

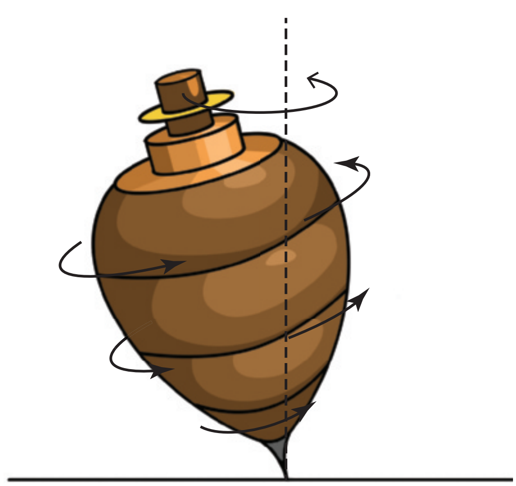


Figure 15.78 A spinning top precesses

Aligning the spin of protons

Charged particles have the property of *spin*, and spinning charges behave like tiny magnets. The spins are usually randomly orientated so they will not produce any net observable magnetic effect. But if these particles are placed in a (strong) magnetic field they can align with the field, although not perfectly. In fact, individual protons (in hydrogen atoms) will rotate around the direction of the magnetic field in a way similar to a child's spinning top rotating around the (vertical) direction of the Earth's gravitational field – see Figure 15.78. This kind of rotation around an axis is called **precession**. The frequency of precession is proportional to the strength of the magnetic field. It is called the **Larmor frequency**.

In NMR, single spinning protons in the nuclei of hydrogen atoms are made to precess when the patient is placed in a very strong uniform magnetic field (typically between 1 T and 3 T). This is known as the *primary* magnetic field. Hydrogen atoms are found throughout the body, especially in water molecules. Such magnetic fields maybe be 50 000 times stronger than the Earth's magnetic field, but they are not known to cause any harmful effects (although there are a few well-understood exceptions). Magnetic fields of this strength can only be produced using very high electric currents circulating in coils of wire. This requires that the coils are at very low temperatures, so that they can become superconducting.

Use of RF signals

The spinning hydrogen nuclei can be made to precess *together, in phase*, by excitation from a resonant radio-frequency (RF) magnetic field. The RF causes resonance because it has the same frequency as the Larmor frequency, which will have a value somewhere within the radio wave section of the electromagnetic spectrum, typically about 60 MHz depending on the strength of the magnetic field.

Because the nuclear magnetic fields of the protons are now precessing in phase with each other, they will generate a rotating magnetic field strong enough to be detected as an oscillating voltage by coils placed around the patient.

After the excitation, the spins relax back to their original distribution at a characteristic rate called the **relaxation rate**. This rate varies with the type of tissue, so that determination of relaxation times leads to information about the type of tissue which, in turn, leads to more detailed and better-quality images.

Use of gradient fields

The description of MRI so far has not explained how it is possible to obtain three-dimensional images of patients. By imposing *gradients* of magnetic field in three perpendicular directions (x, y and z), signals from different parts of the patient can be made to resonate at different frequencies, thus allowing reconstruction of the three-dimensional distribution of protons. A typical MRI arrangement is shown in Figure 15.79.

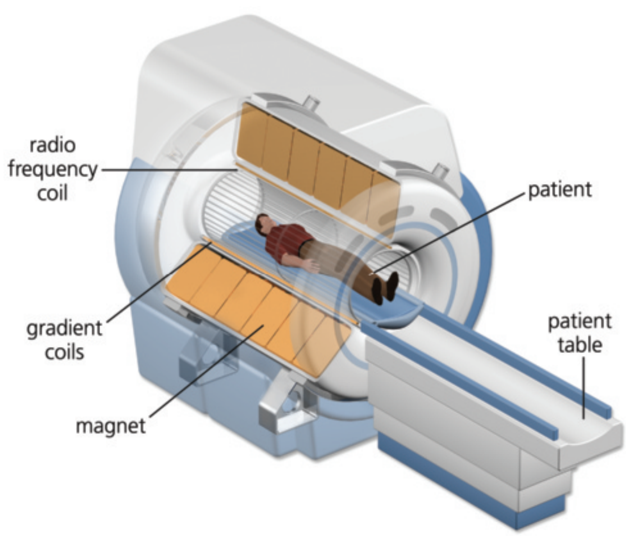


Figure 15.79 Arrangement for magnetic resonance imaging

100 **a** Suggest reasons why some people may find having an MRI scan an unpleasant experience.
b Why would an MRI scan *not* be used to diagnose a broken arm.
c Use the internet to research whether having an MRI scan has any adverse health effects.

101 Explain what is meant by the 'Larmor frequency'.

102 Why are radio frequencies used to excite hydrogen atoms in MRI?

103 Explain what is meant by the term *relaxation time* and why it provides useful information in MRI.

104 **a** What are the essential differences between a CT scan and an MRI scan?
b Why are powerful computers essential for both of these procedures?
c Make a list of the advantages and disadvantages of these two types of scan.

Summary of knowledge

15.1 Introduction to imaging

- When light rays/wavefronts spreading out from a point object are incident on a lens that is thicker in the middle than at its edges, the rays will be refracted and converged to form a real image at the point where the rays cross (unless the object is at the focal point, or nearer to the lens). This kind of lens is called a converging (convex) lens.
- If the rays are incident on a lens that is thinner in the middle than at its edges, the rays will be refracted and diverged and a virtual image can be seen when looking through the lens at the point from which the rays appear to have come. This kind of lens is called a diverging (concave) lens.
- The principal axis of a lens is defined as the imaginary straight line passing through the centre of the lens, which is perpendicular to its surfaces.
- The focal point of a lens is defined as the point through which all rays parallel to the principal axis converge after passing through the lens (or the point from which they appear to diverge).
- The focal length of a lens is defined as the distance between the centre of the lens and the focal point. Its value depends on the refractive index of the material and the curvature of its surfaces.
- The optical power of a lens is defined as 1/focal length; $P = 1/f$. Optical power is measured in dioptres, D. Power (D) = 1/focal length (metres).
- The paths of three rays from the top of any extended object, which pass through a lens and then go to the top of the image, can be predicted. Using these rays, diagrams can be drawn to determine the position and nature of the image formed when objects are placed at various distances from a lens.
- In diagrams and calculations throughout this topic we have assumed that the lens is thin and that the rays are close to the principal axis. If this is not true, the image will not be formed exactly where predicted and the focus/image will not be as well defined.
- Real images are formed where rays actually cross. Virtual images are formed when rays diverge into the eye – the image is formed where the rays appear to have come from.
- The linear magnification of an image is the ratio of the height of the image, h_i, divided by the height of the object, h_o. $m = h_i/h_o = -v/u$.
- The angular magnification of an image is the angle subtended at the eye by the image divided by the angle subtended at the eye by the object. $M = \theta_i/\theta_o$. When referring to optical instruments it is common to refer to angular magnifications rather than linear magnifications.
- If the object is placed further away from a converging lens than the focal point, the image formed will always be real and inverted.
- The thin lens formula is $1/f = 1/v + 1/u$. This formula can be used to determine the position and nature of an image. When using this formula it is important to remember that a distance to a virtual image and the focal length of a diverging lens are always negative. A positive magnification indicates that the image is upright; a negative magnification indicates that the image is inverted ($m = -v/u$).
- If the object is placed at the focal point, or closer to a converging lens, the image will be magnified, upright and virtual. Used in this way the lens is described as a simple magnifying glass.
- The nearest point to the human eye at which an object can be clearly focused (without straining) is called the near point. It is accepted to be 25 cm from a normal eye and often given the symbol D. The furthest point from the human eye that an object can be clearly focused (without straining) is called the far point – for a normal eye it is at infinity.
- The angular magnification, M, of a simple magnifying glass varies between D/f, for the image at infinity, to $(D/f) + 1$ for the image at the near point.

- Lens aberrations (especially with higher-power lenses) are the principal limitations on the magnification achievable by optical instruments that use lenses.
- Spherical aberration produces distorted images. It is the inability of a lens having spherical surfaces to bring all rays incident on it (from a point object) to the same focus. It may be reduced by adapting the shape of the lens, or by only using the centre of the lens.
- Chromatic aberration is the inability of a lens to bring rays of different colours (from a point object) to the same focus. It occurs because refractive index varies slightly with colour (wavelength). It can be reduced by combining lenses of different shapes and refractive indices.
- Diagrams can be drawn to represent these aberrations and how they can be reduced.
- Mirrors with curved surfaces can also be used to focus images. The terminology and the principles involved are very similar to those concerning lenses. Curved mirrors can have spherical aberration problems.

15.2 Imaging instrumentation

- The objective lens of a compound microscope forms a real magnified image of an object that is placed just beyond its focal point. The eyepiece then acts as a magnifying glass to produce a final image, which is inverted, magnified and virtual.
- Ray diagrams can be constructed to represent a microscope in normal adjustment with the final image at (or near to) the near point. Angular magnification equals the linear magnification of objective multiplied by the angular magnification of the eyepiece.
- Resolution is often more important than magnification in optical instruments. Good resolution can be considered to be the ability to see points as being separate. Magnifying an image can improve resolution, but not if the resolution is already poor.
- In general, resolution is improved by using better quality lenses, large apertures and small wavelengths (minimising diffraction effects). Large apertures also have the advantage of collecting more light and producing brighter images.
- Two objects are considered to be *just* resolvable if the angle, θ, that they subtend at the eye or optical instrument is larger than $1.22\lambda/b$ (Rayleigh's criterion), where b is the diameter of the receiving aperture.
- The objective lens of a telescope forms a diminished, real and inverted image of a distant object at its focal point. The eyepiece acts as a magnifying glass to produce a final image at infinity (in normal adjustment), which is inverted, diminished and virtual. The linear magnification is less than one, but the telescope produces an angular magnification, $M = f_o/f_e$.
- Ray diagrams can be constructed to represent a telescope in normal adjustment when the distance between the lenses is $f_o + f_e$.
- Reflecting telescopes use converging mirrors as their objectives (rather than converging lenses).
- Newtonian mountings use plane mirrors to reflect light into an eyepiece lens at the side. Cassegrain mountings use diverging mirrors to increase magnification and enable the observer to look through the telescope directly towards an object.
- Optical astronomical telescopes on the Earth's surface receive light that has been affected by passing through the Earth's atmosphere. Some radiation is absorbed or scattered, and some is refracted irregularly. Placing satellites on orbiting satellites above the atmosphere overcomes these limitations.
- Radio waves (including microwaves) are much less affected by the atmosphere (than light) and radio telescopes can be terrestrial. Many astronomical objects emit radio waves. The simplest radio telescopes have an aerial placed at the focal point of a single parabolic dish reflector.
- As with other waves, resolution is limited to angles larger than $1.22\lambda/b$. Because radio waves from space might have a typical wavelength of about 1 m, good resolution using a single dish can only be achieved if it has a large diameter.
- Higher resolution when receiving radio waves is possible using interferometry techniques, in which the signals from two or more synchronized telescopes are combined electronically

and made to interfere. The spacing and centre of the interference pattern can be used to accurately determine the direction to the source of radiation. The maximum resolution achieved with two telescopes can be calculated using $1.22\lambda/b$, with b equal to their separation. Using many telescopes in an array can improve resolution further.

15.3 Fibre optics

- Most data are sent along cables using either electrical pulses in copper wires, or infrared pulses in optic fibres.
- Data are sent using digital pulses, rather than continuously varying analogue signals. Digital data are transferred as a very large number of pulses, each of which can have only one of two possible levels (commonly called 0 or 1).
- As pulses travel along a cable they attenuate and disperse. Attenuation is the gradual loss of intensity of a signal as it passes through a material. Dispersion is the broadening of the width of a pulse and the associated decrease in intensity.
- These effects limit the distance that data can be transferred (before they need to be amplified and reformed) and the amount of data that can be sent in a given time through a particular cable.
- These effects are significantly lower with infrared pulses in optic fibres than they are with electrical pulses in copper wire.
- Electrical pulses also produce changing electromagnetic fields that can spread away from the cable and cause 'interference' by inducing tiny e.m.f.s in other cables. These random unwanted signals are often called electronic 'noise'. Signals in optic fibres do not have this problem.
- Electronic noise can be significantly reduced in copper cabling by using twisted pairs or co-axial cables.
- Data are transferred in digital form because when pulses are affected by dispersion, attenuation and noise, they can still usually be distinguished as 0s or 1s because there only two distinct levels. (Whereas analogue signals may become too distorted.)
- Optic fibres use the effect known as total internal reflection (Chapter 4). The radiation is reflected internally because the angle of incidence inside the fibre is always larger than the critical angle. In general $n_1/n_2 = 1/\sin c$ or, if air is the external medium: $n = 1/\sin c$.
- The core optic fibre(s) are protected from damage by cladding. The refractive index of the cladding material must be less than that of the core. The cladding also prevents different fibres from coming in contact with each other ('crosstalk').
- Dispersion in optic fibres has two main causes – waveguide dispersion and material dispersion.
- Waveguide dispersion is due to the fact that different rays (that started together) travel along slightly different paths. This problem can be limited by using graded-index fibres, in which the refractive index increases progressively towards the centre. This has the effect of confining rays to curved paths close to the centre of the fibre.
- Step-index fibres have cores of constant refractive index.
- Material dispersion can occur if radiation of different wavelengths is used. This is because they travel at different speeds (so they have slightly different refractive indices). This can be overcome by using monochromatic light (from a laser or infrared LED).
- The intensity of a signal confined to an optic fibre decreases exponentially with distance along the cable. If the intensity decreases from I_0 to I, then the attenuation (in dB) is $10\log_{10} I/I_0$. It is usual to quote an attenuation per unit length of cable (e.g. $-1.5\,\mathrm{dB\,km^{-1}}$). A similar equation can be used for power instead of intensity.
- The decibel (dB) scale is a logarithmic scale commonly used to compare an intensity (or power) to a reference level, especially where there are large differences involved.
- Compared with twisted pair and co-axial cables, optic fibres have lower attenuation, greater data transfer rates, do not produce 'noise', are more secure and are smaller and lighter.

15.4 Medical imaging

- When an X-ray beam is directed at a human body some of the radiation will be absorbed and scattered in the body and some will be transmitted directly, so that some X-rays can be detected on the other side of the body. It is this variation that makes X-rays so useful in medical imaging.
- Different parts of the body will absorb X-rays by different amounts and the intensity of the detected beam will show variations representing the presence of parts of the body with different densities and absorption rates.
- The X-rays that are transmitted can be detected either photographically or by the use of CCDs (charge-coupled devices, as used in digital cameras). The use of CCDs allows the electronic storage and manipulation of images.
- The intensity, I, of a parallel beam of X-rays (not spreading out) decreases exponentially with distance, x, due to absorption and scattering: $I = I_o e^{-\mu x}$. μ is a constant called the linear attenuation coefficient. It represents the amount of attenuation per unit length in a particular medium (for radiation of a specified wavelength). The usual unit is cm^{-1}.
- Attenuation can also be represented in the same way as for attenuation in an optic fibre: attenuation (dB) = $10 \log (I_1/I_0)$.
- Absorption due to the photoelectric effect is the principal means of attenuation of X-rays and it is largely dependent on the proton number, Z, of the atoms present. For example, bone contains elements with a higher average proton number than soft tissue, and therefore absorbs a higher percentage of X-rays.
- The attenuation of X-rays is often characterized by the half-value thickness of a particular medium, $x_{1/2}$, which is defined as the thickness of a medium that will reduce the transmitted intensity to half its previous value.
- Linear attenuation coefficient and half-value thickness are inversely related: $\mu x_{1/2} = \ln 2$.
- The mass attenuation coefficient is used to compare the attenuation in unit masses of different materials: mass attenuation coefficient = linear attenuation coefficient/density = μ/ρ. The usual unit is $cm^2\,g^{-1}$.
- Attenuation of X-rays is greater for lower frequencies (photons of lower energy). Such beams are produced by lower voltages and are often called 'soft' X-rays. 'Hard' X-rays are more penetrating.
- High-quality X-ray images should have high intensity and contrast. The edges of different areas of the image should be sharp and well resolved. But at the same time it is important, for safety reasons, that the power of X-rays used should be as low as possible.
- Techniques for improving the quality of the images include: using a small source that is not too close to the patient; placing an oscillating collimating grid between the patient and the detector; using intensifying screens containing fluorescent materials.
- Computed tomography (CT) uses computer-controlled X-rays and machinery to obtain sharp images of planes of the patient (scans) with good resolution. These scans can be combined to present a three-dimensional image.
- There is a health risk associated with all uses of X-rays. Ultrasound imaging has no known risk, but the images have disappointing resolution because of the relatively long wavelengths used.
- Ultrasound waves (sound with frequencies higher than can be heard by humans) are directed into a patient's body and reflect back off boundaries between different media.
- Acoustic impedance, Z, is a measure of the opposition of a medium to the flow of sound through it. $Z = \rho c$, where ρ is the density of the medium and c is the speed of the wave in the medium. The units of acoustic impedance are $kg\,m^{-2}\,s^{-1}$.
- The bigger the difference in impedances, the higher the percentage of incident waves that are reflected at a boundary between two media.
- Ultrasound waves are produced using the piezoelectric effect, in which an alternating voltage applied across a crystal transducer makes it vibrate at the same frequency, sending waves into

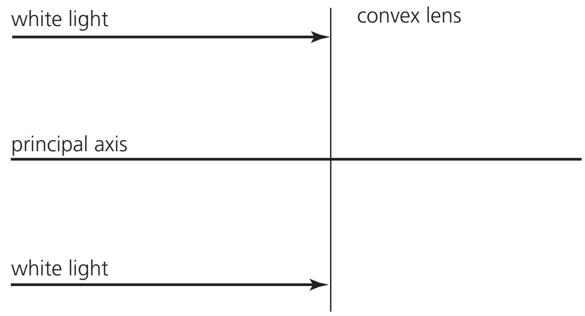
the surroundings. When reflected waves are received back at the probe, oscillating voltages are produced and detected.

- The transducer (probe) is placed next to the skin of the patient with a gel between them (eliminating air). The gel has an acoustic impedance chosen to transmit the waves into the body efficiently.
- The ultrasound waves are transmitted in pulses, with sufficient time between the pulses for the reflected waves to be clearly detected. Resolution is improved by having several complete ultrasound waves in each pulse.
- The simplest types of ultrasound scans are known as A-scans (amplitude scans). The amplitude of the waves reflected from different boundaries in the patient's body are displayed as an amplitude × time graph. Information from the graph can be used to determine the position and size of various parts of the body.
- B-scans are widely used in hospitals. The information is obtained in essentially the same way as in an A-scan, except that the amplitude of the reflected wave is represented by the brightness of a dot on a screen. A two-dimensional real-time video image picture is constructed by a computer programme using the information from one or more transducers inside the ultrasound probe, transmitting waves in a slightly different direction, often while the probe is moved to different positions.
- Higher ultrasound frequencies (smaller wavelengths) have less diffraction so that the beams are more directional and the images have better resolution. However, higher frequencies also undergo more attenuation. The frequency being used may need to be changed depending on the particular circumstances.
- Despite its poor resolution, ultrasound imaging provides a quick, safe, economical and mobile way of examining inside the body, especially when soft tissues are involved. Ultrasound cannot penetrate into bone effectively and cannot be used for spaces that contain air (e.g. lungs).
- Nuclear magnetic resonance (NMR) provides an alternative to CT scans for providing images of sections through the body. Medical applications of NMR are also commonly called magnetic resonance imaging, MRI. Because MRI does not involve ionizing radiation, it is considered safer than X-ray processes.
- MRI uses the spins of protons in hydrogen atoms. Hydrogen atoms are found throughout the body, particularly in water molecules.
- Resonance is the name given to the effect in which a system (that can oscillate) absorbs energy from another external oscillating source.
- Protons spin and behave like tiny magnets. These spins are usually randomly orientated so that they will not produce any net observable magnetic effect. During an MRI scan the patient is placed in a very strong primary magnetic field. This causes the spinning protons to precess around the direction of the external field.
- The rate of precession is called the Larmor frequency and it is proportional to the strength of the applied magnetic field. The Larmor frequency is in the radio-wave (RF) section of the electromagnetic spectrum.
- When protons are also subjected to an oscillating electromagnetic field of the same frequency (provided by the RF coils), resonance occurs and the protons begin to spin together, in phase. This affects the overall magnetic field strength.
- After the external RF signal is stopped the protons return to their earlier state at a rate that depends on the kind of tissue in which they are located. The changing magnetic field can be detected by the RF coils. The different relaxation rates provide information about the type of tissue.
- As well as the primary magnetic field, by imposing gradients of magnetic field in three perpendicular directions (x, y and z), signals from different parts of a patient can be made to resonate at different frequencies, allowing reconstruction of the three-dimensional distribution of protons.

Examination questions – a selection

Paper 3 IB questions and IB style questions

Q1 a Two parallel rays of white light are incident on a convex lens.



On a copy of the diagram, after refraction in the lens, draw the paths for the rays of red light and blue light present in the white light. (2)

b Use your diagram in **a** to explain chromatic aberration. (3)

c State one way in which chromatic aberration may be reduced. (1)

d An object is placed 5.0 cm from the lens and is illuminated with red light. The focal length of the lens for red light is 8.0 cm. Calculate the:

i position of the image (2)

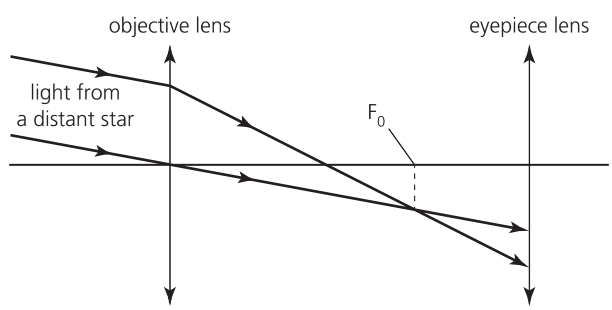
ii linear magnification. (1)

Q2 a Draw a ray diagram to show how a converging mirror can be used to produce an inverted, magnified, real image. (3)

b i Describe the images formed by diverging mirrors. (2)

ii Give one everyday use for diverging mirrors. (1)

Q3 a The diagram shows two rays of light from a distant star incident on the objective of an astronomical telescope. The paths of the rays are also shown after they pass through the objective lens and are incident on the eyepiece lens of the telescope.



The principal focus of the objective lens is F_o. On a copy of the diagram, mark the position of the:

i principal focus of the eyepiece lens (label this F_e) (1)

ii image of the star formed by the objective lens (label this I). (1)

b State where the final image is formed when the telescope is in normal adjustment. (1)

c Complete the diagram in **a** to show the direction in which the final image of the star is formed for the telescope in normal adjustment. (2)

d The eye ring of an astronomical telescope is a device that is placed outside the eyepiece lens of the telescope at the position where the image of the objective lens is formed by the eyepiece lens. The diameter of the eye ring is the same as the diameter of the image of the objective lens. This ensures that all the light passing through the telescope passes through the eye ring.
A particular astronomical telescope has an objective lens of focal length 98.0 cm and an eyepiece lens of focal length 2.00 cm (i.e. f_o = 98.0 cm, f_e = 2.00 cm). Determine the position of the eye ring. (4)

Q4 **a** Explain what is meant by the *angular magnification* produced by an optical instrument. (2)
b A small object is viewed through a converging lens of focal length 6.8 cm used by a student as a magnifying glass. If the image is at infinity what is the angular magnification achieved with normal vision? (2)
c In order to obtain a greater magnification the student constructs a compound microscope that has two lenses of powers 100 D and 25 D.
i What are the focal lengths of these two lenses? (1)
ii Which of these two lenses is used as the eyepiece of the microscope? (1)
iii What is the linear magnification provided by the objective lens when an object is placed 1.2 cm in front of it? (2)
iv Calculate the angular magnification achieved by the microscope when in normal adjustment, with the image at the near point. (2)

Q5 **a** A signal of power 53 mW enters an optic fibre of length 10.4 km. If the power of the signal at the end of the cable has reduced to 32 mW, calculate the attenuation per km along the cable. (2)
b Dispersion is a cause of attenuation in the cable. Distinguish between waveguide dispersion and material dispersion. (3)
c Explain how the use of graded-index fibres reduces waveguide dispersion. (2)

Q6 **a** Outline how digital data are transferred using coaxial cables. (2)
b List two advantages of the use of optic fibres for transferring data, compared with coaxial cables. (2)

Higher Level only

Q7 **a** Define half-value thickness. (1)
b The half-value thickness in tissue for X-rays of a specific energy is 3.50 mm. Determine the fraction of the incident intensity of X-rays that has been transmitted through tissue of thickness 6.00 mm. (3)
c For X-rays of higher energy than those in **b**, the half-value thickness is greater than 3.50 mm. State and explain the effect, if any, of this change on your answer in **b**. (2)
d X-ray images are often blurred despite the patient remaining stationary during exposure.
i State one possible physical mechanism for the blurring of an X-ray image. (1)
ii For the physical mechanism stated in **d i** suggest how X-ray images can be made more distinct. (2)
e The exposure time of photographic film to X-rays is longer than that for visible light. The exposure time for X-rays may be reduced with the use of enhancement techniques, such as that of an intensifying screen. Outline how an intensifying screen reduces the exposure time. (2)

Q8 **a** Computed tomography (CT) scans can provide much more useful information than individual X-rays. Outline the techniques used in CT scans that produce this improvement. (3)
b **i** Give two advantages that CT scans have compared with ultrasound scans. (2)
ii Give two advantages that ultrasound scans have compared with CT scans. (2)

Q9 **a** State the approximate range of ultrasound frequencies used in medical imaging. (1)

b Distinguish between an A-scan and a B-scan. (1)

c State one advantage and one disadvantage of using ultrasound at a frequency in the upper part of the range stated in **a**. (2)

d A parallel beam of X-rays of a particular energy is used to examine a bone. At this energy the half-value thickness of bone is 0.012 m and of muscle is 0.040 m. The beam passes through bone of thickness 0.060 m and through muscle of thickness 0.080 m. Determine the ratio:

$$\frac{\text{decrease in intensity of beam produced by bone}}{\text{decrease in intensity of beam produced by muscle}}$$ (3)

e Suggest, using your answer to **d**, why is this beam is suitable for identifying a bone fracture. (1)

Q10 Nuclear magnetic resonance (NMR) provides an alternative to CT scans for obtaining images from inside the body.

a Explain in general terms what is meant by *resonance*. (2)

b During an NMR scan, what parts of the patient are made to resonate, and how is this resonance produced? (3)

c Why is NMR usually considered to be safer than the use of X-rays? (1)

Option D

16 Astrophysics

ESSENTIAL IDEAS

- One of the most difficult problems in astronomy is coming to terms with the vast distances between stars and galaxies and devising accurate methods for measuring them.
- A simple diagram that plots the luminosity versus the surface temperature of stars reveals unusually detailed patterns that help understand the inner workings of stars. Stars follow well-defined patterns from the moment they are created to their eventual death.
- The Hot Big Bang model is a theory that describes the origin and expansion of the universe and is supported by extensive experimental evidence.
- The laws of nuclear physics applied to nuclear fusion processes inside stars determine the production of all the elements up to iron.
- The modern field of cosmology uses advanced experimental and observational techniques to collect data with an unprecedented degree of precision, and as a result very surprising and detailed conclusions about the structure of the universe have been reached.

16.1 (D1: Core) Stellar quantities *– one of the most difficult problems in astronomy is coming to terms with the vast distances between stars and galaxies and devising accurate methods for measuring them*

Nature of Science

A topic without practical investigations

Astronomy is an unusual topic within the study of physics because the standard 'scientific method' is not so obvious. There are no controlled experiments designed to investigate a theory. Instead astronomers make observations and collect data. One consequence of this is that the growth of knowledge in astronomy is very much dependent on the latest technology available to aid observations.

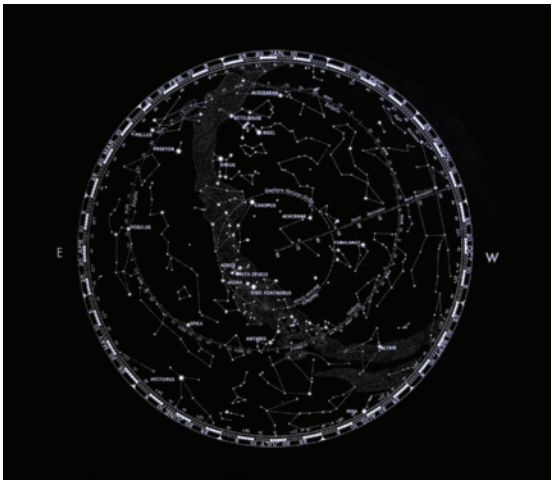
No student studying astronomy can fail to be impressed by the depth of knowledge about the universe that astronomers have gained from apparently so little evidence – just the radiation received from outer space!

In the first section of this chapter we will begin by summarising what we can see in the night sky and then outline the essential features of stars and stellar systems, before explaining the scale of the universe and the units astronomers use to measure such large distances. Finally we will establish the important relationship between the power emitted from a star and the intensity received here on Earth.

Observing the night sky

On a clear night, far away from the light pollution of towns, it may be possible to see hundreds of stars in the night sky with the unaided eye. A total of about 5000 stars are visible from Earth with the human eye, but not all can be seen at the same time, or from the same place. What we can see depends on our location, the time of night and the time of year. This variation happens because of the Earth's motion – its spin on its axis and its orbit around the Sun. In theory, at any one time, in any one place, we might be able to see about half of the visible stars.

Stars *seem* to stay in exactly the same positions/patterns (relative to other stars) over thousands of years and therefore we can locate the stars precisely on a **star map**, such as shown in Figure 16.1. Although stars are moving very fast, their motion is not usually noticeable from Earth, even over very long periods of time (in human terms) because they are such enormous distances away from us.

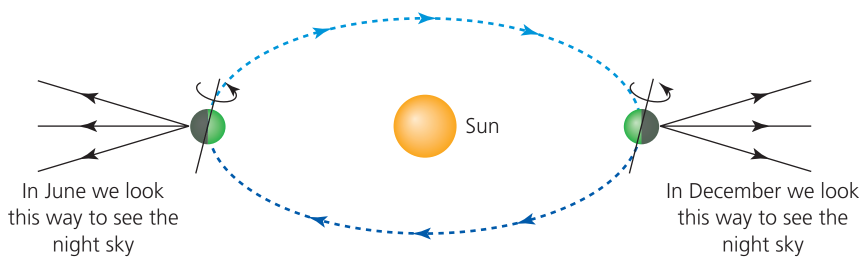


■ **Figure 16.1** A star map for the southern hemisphere

■ **Figure 16.2** The apparent rotation of the stars as the Earth spins

If you observe the stars over a period of hours on any one night you will notice that they appear to move across the sky from east to west – in exactly the same way as the Sun appears to move during the day. These apparent motions are actually produced because the Earth spins in the opposite direction. Time-lapse photography can be used to show the paths of stars across the sky during the night. Such photographs can even show the complete circular path of stars which are close to the Earth's extended axis (Figure 16.2).

In the course of one day, the Earth's rotation causes our view of the stars to revolve through 360° but, of course, during the day we are not able to see the stars because of the light from the Sun. (Radio astronomers do not have this problem.) Our night-time view changes slightly from one night to the next and after six months we are looking in exactly the opposite direction, as shown in Figure 16.3. The Sun, the Moon and the five planets that are visible with the unaided eye are all much, much closer to Earth than the stars. Their movements as seen from Earth can seem more complicated and they cannot be located in fixed positions on a star map. The Sun, the Earth, the Moon and the planets all move in approximately the same plane. This means that they follow similar paths across the sky as seen by us as the Earth rotates.



■ **Figure 16.3** How our view of the night sky changes during the year

The Sun and the Moon are the biggest and brightest objects in the sky. In comparison, all stars appear only as points of light. The closest planets may just appear as discs (rather than points) of light, especially Venus which is the brightest natural object in the night sky (other than the Moon).

There are a few other things we might see in the night sky. At certain times, if we are lucky, we may also be able to see a **comet**, an *artificial satellite* or a *meteor* – which causes the streak of light seen in the sky when a rock fragment enters the Earth's atmosphere and burns up due to friction. Occasionally, parts of meteors are not completely vaporized and they reach the Earth's surface. They are then called *meteorites* and are extremely valuable for scientific research, being a source of extra-terrestrial material.

Astrophysics internet sites

For many students astrophysics is a fascinating subject, but the opportunities for practical work are obviously limited. However, a considerable amount of very interesting information and stunning images are available on the internet and, without doubt, it will greatly enhance the study of this topic if you have easy and frequent access to the websites of prominent space organizations such as the European Space Agency (ESA), NASA and Hubble among others.

Objects in the universe

In this option we will concentrate our attention on (in order of size):

- planets and planetary systems (like the solar system), including comets
- stars (single and binary), stellar clusters (open and globular) and constellations
- nebulae
- galaxies, clusters of galaxies and super clusters of galaxies.

Nebulae

Nebulae are enormous diffuse 'clouds' of **interstellar matter**, mainly gases (mostly hydrogen and helium) and dust. Some of the matter may be ionized. A nebula forms over a very long time because of the gravitational attraction between the masses involved. ('Interstellar' means between the stars.)

There are several kinds of nebulae, with different origins and different sizes. Large nebulae are the principal location for the formation of stars and most nebulae already contain stars that are the source of the energy and light by which they can be observed.

It is possible to see some nebulae in our galaxy without a telescope, although they are diffuse and dim. They were probably first observed nearly 2000 years ago. Recent images of nebulae taken from the Hubble telescopes are truly spectacular. Figure 16.4 shows a telescope image of the Orion nebula. This can be seen without a telescope (close to Orion's belt in the Orion constellation) and it contains a number of 'young' stars. It is one of the closest nebulae to Earth and one of the brightest, so it has been much studied as a source of information about the formation of stars. It is about 1×10^{16} km from Earth and about 2×10^{14} km in diameter, so that it subtends an angle at the eye of approximately 0.02 rad (≈ 1°, which is large in terms of astronomy).

Figure 16.4 The Orion nebula (as pictured through a telescope)

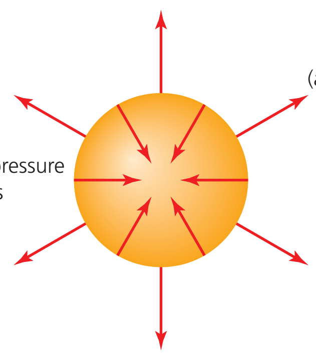
Stars

Within part of a nebula, over a very long period of time, gravity pulls atoms closer together and they can gain very high kinetic energies (that is, the temperature is extremely high – millions of kelvin) if the overall mass is large. The hydrogen nuclei (protons) can then have enough kinetic energy to overcome the very high electric forces of repulsion between them and fuse together to make helium nuclei. This process, known as **nuclear fusion**, can be simplified to:

$$4\,{}^{1}_{1}\text{H} \rightarrow {}^{4}_{2}\text{He} + 2\,{}^{0}_{1}\text{e} + \text{neutrinos and photons}$$

> Nuclear fusion happens in all stars (until near the end of their 'lifetimes') and is their dominant energy transformation.

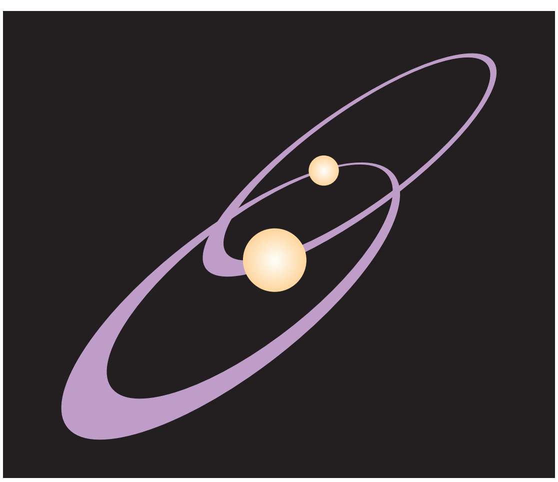
Each completed nuclear fusion of helium from four hydrogen nuclei (protons) is accompanied by a decrease in mass and an equivalent release of energy amounting to about 27 MeV (Chapter 12). The fusion of heavier elements occurs later in the lifetime of stars.



■ **Figure 16.5** A stable star in equilibrium

When nuclear fusion begins on a large scale it is commonly described as the *birth* of a star. The contraction of the material in the forming star creates a **thermal gas pressure** and the emitted radiation also creates a **radiation pressure** outwards in opposition to the **gravitational pressure** inwards. These pressures remain equal and opposite for a very long time, during which the star will remain the same size, stable and unchanging. It will be in **stellar equilibrium** (Figure 16.5). It may be helpful to compare this to a balloon in equilibrium under the action of the gas pressure outwards and the pull of the elastic inwards. There is also a balance between energy transferred from fusions and energy radiated from the surface.

During this period the star is known as a **main sequence** star. The only fundamental difference between these stars is their masses. Eventually the supply of hydrogen will be used up and the star will no longer be in equilibrium. This will be the beginning of the end of the 'lifetime' of a main sequence star. What happens then depends on the mass of the star (explained later in this chapter). Our Sun is approximately halfway through its lifetime as a main sequence star.



■ **Figure 16.6** An artist's impression of a visual binary star system

Binary stars

It is estimated that around half of all stars are in fact two (or more) stars orbiting around their common centre of mass with a constant period. Stars in a two-star system are described as **binary stars** (see Figure 16.6). Binary stars that are not too far away from Earth may be seen through a telescope as two separate stars, but most binary stars are further away and appear as a single point of light.

Binary star systems are important in astronomy because the period of their orbital motion is directly related to their mass. This means that if we can measure their period, we can calculate their mass. For non-visual binaries this may be possible using one of two observations:

- If one star passes regularly in front of the other as seen from Earth (an *eclipse*), the brightness will change periodically.
- If one star is momentarily moving towards the Earth, the other must be moving in the opposite direction. The frequency of the light received on Earth from each will be Doppler-shifted (Chapter 9) periodically.

Groups of stars

Galaxies

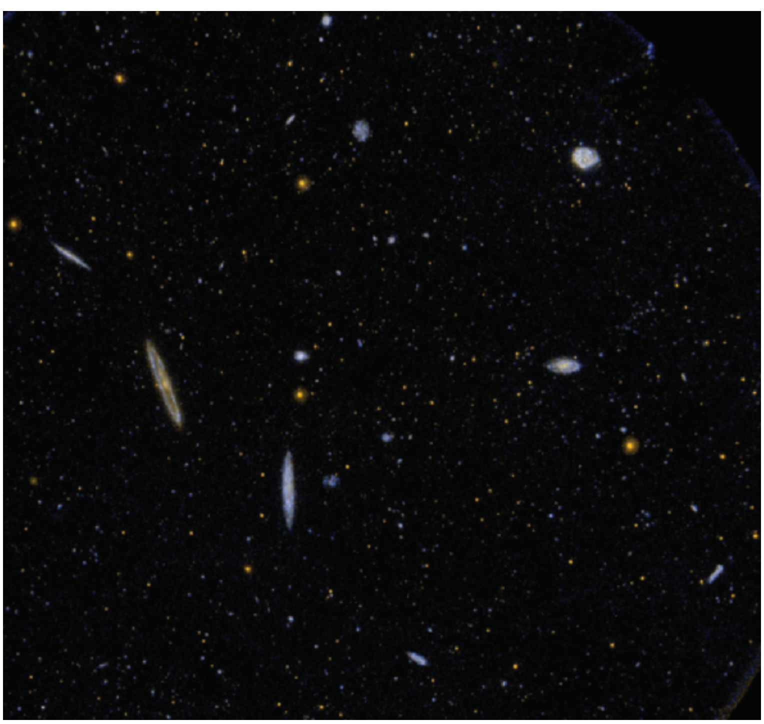
When we look at the stars in the night sky, they seem to be distributed almost randomly, but we are only looking at a tiny part of an enormous universe. The force of gravity causes billions of stars to collect into groups, all orbiting the same centre of mass. These groups are known as **galaxies**. Some of the spots of light we see in the night sky are distant galaxies (rather than individual stars). Billions of galaxies have been observed using astronomical telescopes. The Earth, the Sun and all the other stars that we can see with the unaided eye are in a galaxy called the **Milky Way**.

■ **Figure 16.7** Spiral galaxy M81

Galaxies are commonly described by their shape as being *spiral* (Figure 16.7), *elliptical* or *irregular*.

Galaxies are distributed throughout space, but not in a completely random way. For example, the Milky Way is one of a group of about 50 galaxies known as the 'Local group'. Larger groups of galaxies, called **clusters of galaxies**, are bound together by gravitational forces. (See Figure 16.8 for an example.) Clusters may contain thousands of galaxies and much intergalactic gas along with undetected 'dark matter'. (The term 'galactic cluster' is commonly used for a group of *stars* within a galaxy.)

Clusters of galaxies are not distributed evenly throughout space, but are themselves grouped together in what are known as **super clusters**. Super clusters of galaxies may be the largest 'structures' in the universe.



■ **Figure 16.8** Virgo cluster of galaxies

Stellar clusters

Some stars within a galaxy are close enough to each other that they become gravitationally bound together and rather than move independently, they move as a group called a **stellar cluster**. All the stars within a particular cluster were formed from the same nebula. There are two principal types of stellar cluster:

- **Globular clusters** are old and contain many thousands of stars in roughly spherical shapes that are typically about 10^{14} km in diameter.
- **Open clusters** are not as old as globular clusters. They are about the same size but contain much fewer stars (typically a few hundred). Because there are fewer stars in an open cluster, the overall shape is less well defined and the gravitational forces are weaker. Over time, an open stellar cluster may disperse because of the effects of other gravitational forces. The Pleiades (Figure 16.9) are an open cluster that is visible from Earth without a telescope.

■ **Figure 16.9** The Pleiades are an open stellar cluster.

It is important not to confuse stellar clusters, which are groups of stars relatively close to each other in space, with constellations.

Constellations

Ancient societies, such as Chinese, Indian and Greek civilizations, attempted to see some order in the apparent random scattering of the stars that we can see from Earth. They identified different parts of the night sky by distinguishing patterns of stars representing some aspect of their culture, such as the Greek hunter Orion (see Figure 16.10).

■ **Figure 16.10** The constellation of Orion: (a) the stars seen in the sky, (b) a representation from mythology

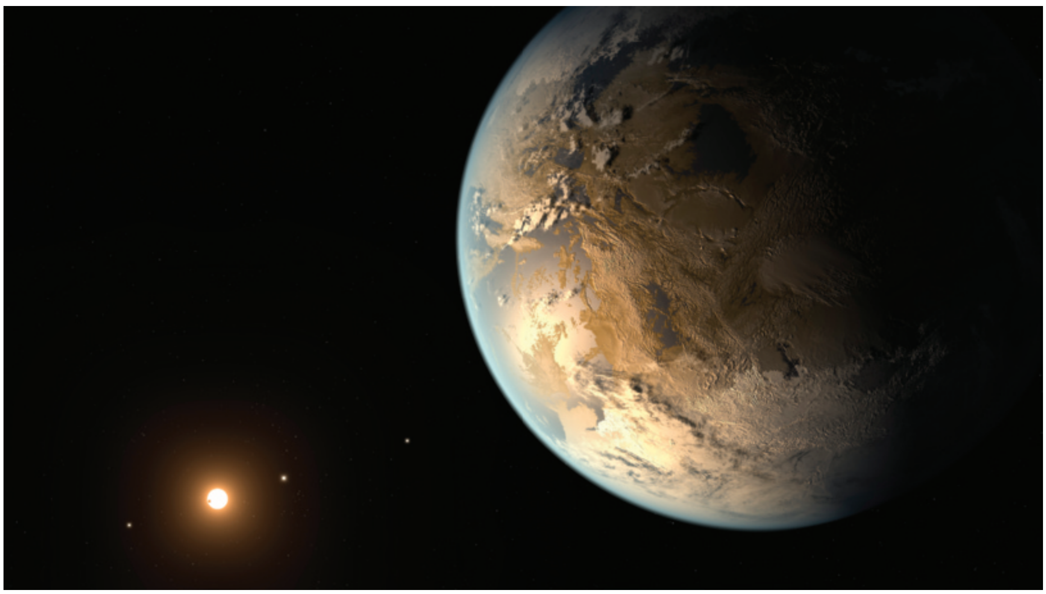
These two-dimensional patterns of visible stars are called **constellations**. It is important to understand that the stars within any given constellation do not necessarily have anything in common. They may not even be 'close together', despite the impression we have by viewing them from Earth. Although many constellations were first named thousands of years ago, their names are still widely used today to identify parts of the night sky.

Planetary systems around stars

A **planetary system** is a collection of (non-stellar) masses orbiting a star. Planetary systems are believed to be formed by the same processes as the formation of the stars.

Planets are objects of sufficient mass that gravitational forces have formed them into spherical shapes, but their mass is not large enough for nuclear fusion to occur. In other words, they are not massive enough to be stars. To distinguish planets from some smaller orbiting masses, it has been necessary for astronomers to specify that a planet has 'cleared its neighbourhood' of smaller masses close to its orbit.

The search for extraterrestrial intelligence (SETI) concentrates on planetary systems like our own solar system and new planetary systems are now discovered regularly. By the start of the year 2014 more than one thousand planetary systems had been identified. In April 2014 astronomers announced that they had discovered the 'most Earth-like planet', Kepler 186f, orbiting a small star at a distance of about 500 ly (light years) from Earth (see Figure 16.11).



■ **Figure 16.11** An artist's impression of Kepler 186f

The solar system

The **Sun** and all the objects orbiting it are collectively known as the **solar system**. Our Sun is a star and it is very similar to billions of other stars in the universe. It has many objects orbiting around it that are held in their orbits by gravity. The solar system is an example of a planetary system. Most of the planets have one or more objects orbiting around them. These are called **moons**. The Sun is the only large-scale object in our solar system which emits visible light; the others are only visible because they reflect the Sun's radiation towards Earth.

The Sun was formed about 4.6 billion years ago from the collapse of an enormous cloud of gas and dust. Evidence from radioisotopes in the Earth's surface suggests that the Earth was formed about the same time, 4.5 billion years ago.

Table 16.1 shows some details of the planets of our solar system (which do *not* need to be remembered). The distances given in the table are only averages because the planets are not perfect spheres and because their orbits are **elliptical** (oval) rather than circular. The Earth's orbit, however, is very close to being circular so the Earth is always about the same distance from the Sun. (The Earth is closest to the Sun in January but there is only about a 3% difference between the smallest and largest separations.) An ellipse has two *focuses* (foci) and the Sun is located at one of those two points. The **period** of the Earth's orbit is, of course, one year, but note that the further a planet is from the Sun, the longer its period. The link between orbital radius and period was discussed in Chapters 6 and 10.

Table 16.1 Planetary data (all data is correct to two significant figures)

Planet	Mass/10^{24} kg	Radius of planet/10^6 m	Mean distance from Sun/10^{11} m	Period/y
Mercury	0.33	2.4	0.58	0.24
Venus	4.9	6.1	1.1	0.62
Earth	6.0	6.4	1.5	1.0
Mars	0.64	3.4	2.3	1.9
Jupiter	1900	69	7.8	12
Saturn	570	57	14	29
Uranus	87	25	29	84
Neptune	100	25	45	160

Compared with planets, **comets** are relatively small lumps of rock and ice that also orbit the Sun, but typically with very long periods and very elliptical paths (see Figure 16.12). They therefore spend relatively little of their time in the inner solar system close to the Sun and the inner planets, such as Earth. When they approach the Sun, radiation and the outflow of particles (solar wind) often cause a comet to develop a diffuse tail of dust and gas, which always points away from the Sun (Figure 16.13). This, together with the rarity of seeing them, has made comets a matter of great curiosity for many of the world's civilizations. Probably the most famous comet is named after the British astronomer and mathematician Edmund Halley (1656–1742). Halley correctly predicted that this comet would next be seen in 1758 (which was 16 years after his death). Halley's comet has a period of 75 years; it was last seen in 1986 and will be seen next in the year 2061.

In November 2014, after a 10-year mission, the European Space Agency's spacecraft *Rosetta* landed the first object on a comet. The Philae lander was able to identify organic molecules on comet 67P.

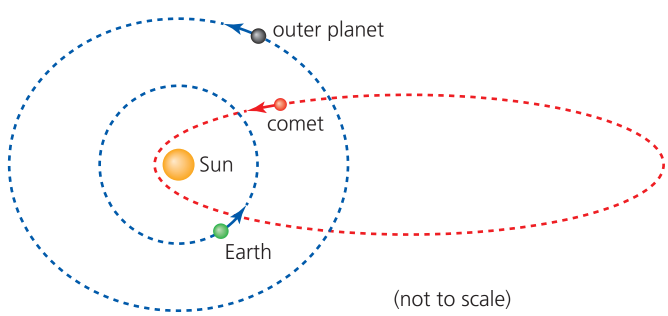


Figure 16.12 The eccentric ('flattened') path of a comet

Figure 16.13 A comet and its tail

1 a Calculate the average density of Earth and Jupiter.
  b Why are they so different?

2 a What is the average orbital speed of the Earth?
  b Compare the Earth's speed to that of Mercury.

3 a If there was a planet located at 35×10^{11} m from the Sun, suggest how long it might take to complete its orbit.
  b Would such a planet be visible to the unaided eye? Explain your answer.

4 a What is the smallest planet and what is its mass?
  b Why is Pluto not considered to be a planet?

5 What is the largest planet and what is its diameter?

Additional Perspectives

Asteroids colliding with the Earth

Asteroids are large rocks that are generally bigger than comets but much smaller than planets. They do not have 'tails' and most orbit the Sun in approximately circular orbits between Mars and Jupiter, in a zone called the *asteroid belt*. Because they are relatively small, the *trajectories* (paths) of asteroids and comets may be significantly altered if they pass 'close' to a planet (especially Jupiter) when they are subject to large gravitational forces.

Science-fiction authors and movie makers enjoy frightening us all with stories about asteroids or comets colliding with the Earth, but it is only in recent years that scientists have come to realise that such a major collision is not as unlikely as they had previously thought. In 1994 a large comet (Shoemaker–Levy 9) collided with Jupiter. The effect of the impact was seen easily with a telescope and was broadcast around the world on television (Figure 16.14). If a similar comet collided with Earth, the results would be catastrophic, although not quite on a scale comparable to the asteroid collision with Earth about 65 million years ago, which is thought to have led to the extinction of many species, including the dinosaurs.

Figure 16.14 Astronomers watch the impact of comet Shoemaker–Levy 9 with Jupiter.

We only need to look at the crater-covered surface of the Moon to become aware of the effects of collisions with asteroids and comets, but similar evidence is not so easy to find on the Earth's surface. Rocks of diameter 10 m or less usually break up in the Earth's atmosphere before impacting, so an asteroid would need to have a diameter of about 50 m or more before its impact would leave a noticeable and long-lasting crater. The effects of friction with the air might also cause an asteroid to explode before it impacted the Earth's surface. Of course, most of the Earth is covered with water and no craters would be formed after an impact with the oceans. Also, old craters may well have been eroded, weathered or just covered with vegetation over long periods of time.

Actual estimates about the size of possible asteroids that could collide with Earth and the probability of such events occurring are continually being refined. But, in general terms, we know that the probability of the Earth being struck by an asteroid is inversely related to its size.

For example, an asteroid 50 m in diameter may impact the Earth about every 1000 years; a 1 km asteroid about every 500 000 years and a 10 km asteroid once every 100 000 000 years. The chance of a catastrophic impact in an average human lifetime may be about 1 in 10 000.

There may be up to a million asteroids in our solar system capable of destroying civilization if they impacted with Earth, but it is not easy to observe all of them, nor track their movements. Much effort is now going into Near Earth Objects programs and researching what might be done if a dangerous impact was expected.

1 Calculate the kinetic energy of an asteroid of diameter 1 km and average density of 400 $kg\,m^{-3}$ travelling at a speed of 20 $km\,s^{-1}$. Compare your answer with 25 megatonnes of TNT, the energy that would be released from a 'large' nuclear bomb. (1 tonne of TNT is equivalent to 4.2×10^9 J.)

2 Use the internet to find out when the next large asteroid is expected to pass near to Earth. How close will it come and how dangerous would it be if it hit us?

Astronomical distances

The universe is enormous! Rather than use metres (or km) to measure distances, astronomers usually prefer to deal with smaller numbers and have introduced alternative units for distance.

The **light year**, **ly**, is defined as the *distance* travelled by light in a vacuum in one year.

At a light speed of $2.998 \times 10^8\,\text{m s}^{-1}$ and 365.25 days, a light year is easily shown to be 9.46×10^{15} m. This value is provided in the *Physics data booklet*.

The **astronomical unit**, **AU**, is equivalent to the mean distance between the Earth and the Sun, 1.50×10^{11} m.

This value is provided in the *Physics data booklet*. (Although the actual distance varies, the value of 1 AU is *defined* to be $1.495\,978\,707 \times 10^{11}$ m.)

One **parsec**, **pc**, is equal to 3.26 ly. This value is provided in the *Physics data booklet*. The parsec is the preferred unit of measurement in astronomy because it is closely related to parallax angles – the way in which the distances to 'nearby' stars are measured (this will be explained later).

One parsec is defined as the distance to a star that has a parallax angle of one arc-second.

While distances to 'nearby' stars are commonly measured in parsecs, the more distant stars in a galaxy are kpc away and distances to the most distant galaxies will be recorded in Mpc and Gpc.

Table 16.2 Summary of distance units commonly used in astronomy

Unit	Metres/m	Astronomical units/AU	Light years/ly
1 AU =	1.50×10^{11}	–	–
1 ly =	9.46×10^{15}	6.30×10^4	–
1 pc =	3.09×10^{16}	2.06×10^5	3.26

The scale of the universe

The diameter of the **observable universe** is about 9×10^{10} ly. The speed of light limits the amount of the universe that we can, in principle, 'observe'. The distance to the edge of the observable universe is equal to the speed of light multiplied by the age of the universe (but the expansion of space itself must be considered, which will be discussed later in the chapter).

Distances between stars and between galaxies vary considerably. As a *very approximate* guide there might be 10^{12} stars in a big galaxy. A typical separation of stars within it may be about 1 ly, with a typical total diameter of a galaxy being about 10^4 ly (Figure 16.15). The billions of galaxies are separated from each other by vast distances, maybe 10^7 ly or more.

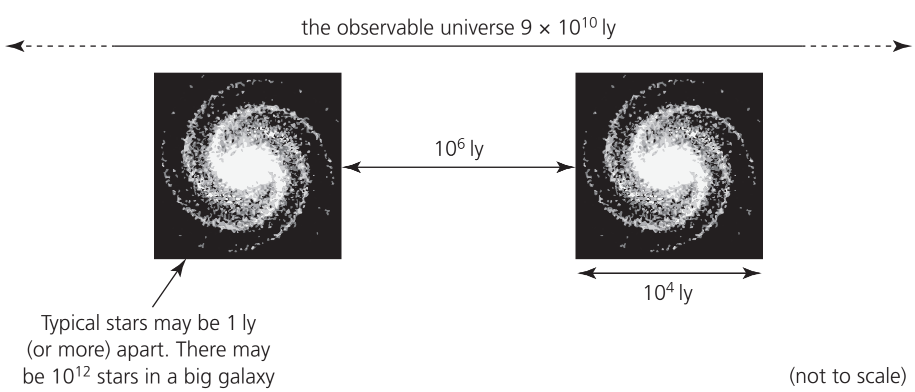


Figure 16.15 Very approximate dimensions of galaxies

6 Use Table 16.1 to determine the mean distance (in AU) from the Sun to the planets Mercury and Uranus.

7 What is the approximate size of the observable universe in:
  a km
  b pc?

8 Proxima Centauri is the nearest star to Earth at a distance of 4.0×10^{16} m.
  a How many light years is this?
  b If the Earth was scaled down from a diameter of 1.3×10^{7} m to the size of a pin head (1 mm diameter), how far away would this star be on the same scale?

9 Our solar system has an approximate size of at least 10^{11} km.
  a How many light years is that?
  b If you were making a model of our solar system using a ball of diameter 10 cm to represent the Sun, how far away would the 'edge' of the solar system be? (The Sun's diameter = 1.4×10^{6} km.)
  c Research into how the edge of the solar system can be defined and what objects in the solar system are the most distant from the Sun.

10 Calculate the time for light to reach Earth from the Sun.

11 a Estimate how long would it take a spacecraft travelling away from Earth at an average speed of $4\,\text{km s}^{-1}$ to reach:
    i Mars
    ii Proxima Centauri.
  b Find out the highest recorded speed of a spacecraft.

12 Use the data from Figure 16.15 to make a very rough estimate of the number of stars in the observable universe.

13 Research the diameter of our galaxy, the Milky Way, in parsecs.

14 Explain why it would be unusual to quote a distance between stars in AU.

ToK Link

Imagination

The vast distances between stars and galaxies are difficult to comprehend or imagine. Are other ways of knowing more useful than imagination for gaining knowledge in astronomy?

Imagining the vast distances in the universe may be considered to be similar to imagining the number of molecules in a grain of salt – the numbers are so large that they are almost meaningless to us. There is no doubt that it does help us to make comparisons like 'it would take more than a billion years to walk to the nearest star', but then we realise that this is an incredibly *small* distance in the universe!

Determining the distances to the stars and distant galaxies

The measurement of astronomical distances is a key issue in the study of astronomy. However, determining the distance from Earth to a star or galaxy accurately is not easy and a variety of methods have been developed.

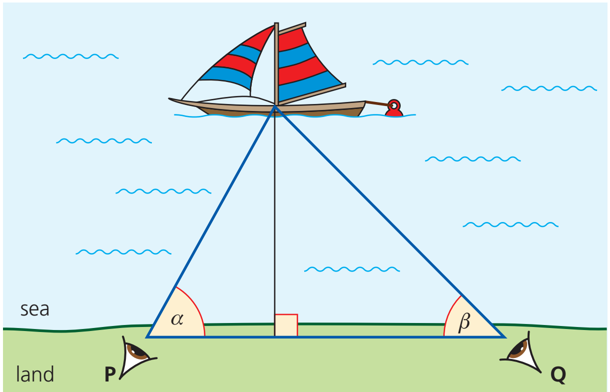
In this course we will consider three different ways in which the distance to a star or distant galaxy may be determined:

- stellar parallax
- use of Cepheid variable stars
- use of supernovae.

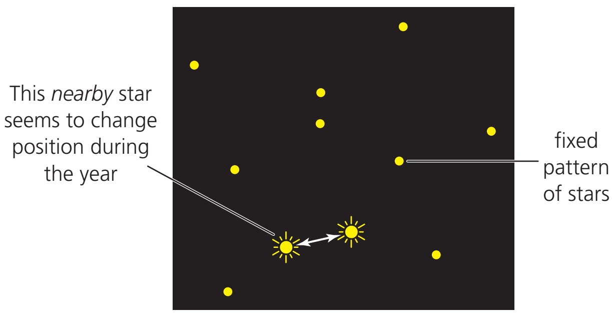
The use of **stellar parallax** for 'nearby' stars is the most direct and easily understood method. The other two methods are used for distant galaxies. They will be discussed later in the chapter.

Stellar parallax and its limitations

This method is similar in principle to one that we might use on Earth to determine the distance to an inaccessible object, such as a boat or a plane. If the object can be observed from two different places, then its distance away can be calculated using trigonometry. An example of this *triangulation* method is shown in Figure 16.16.



■ **Figure 16.16** Determining the distance to a ship at sea using triangulation



■ **Figure 16.17** A nearby star's apparent movement due to parallax

An observer on land sees the boat from position P and then moves to position Q. If the angles α and β are measured and the distance PQ is known, then the other distances can be calculated. When astronomers want to locate a star, they can try to observe it from two different places, but the distance between two different locations on Earth is far too small compared with the distance between the Earth and the star. Therefore, astronomers observe the star from the same telescope at the same location, but at two different places in the Earth's orbit; in other words, at different times of the year. To get the longest difference they usually take two measurements separated in time by six months.

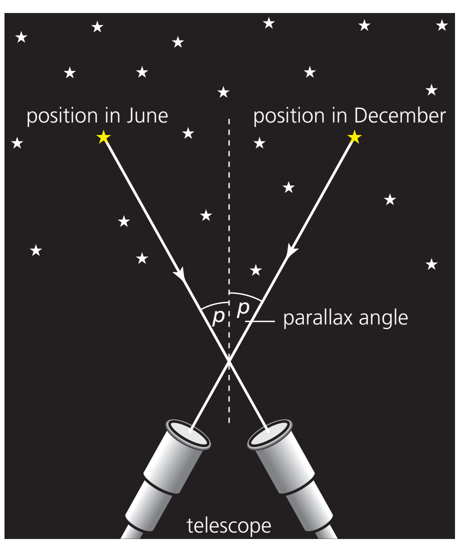
The triangulation method described above to locate a boat would be much more difficult if the observer was in a *moving* boat at sea and this is similar to the difficulty faced by astronomers locating stars from Earth. The problem can be overcome by comparing the position of the star to other stars much further away (in the 'background'). This is known as a **parallax** method.

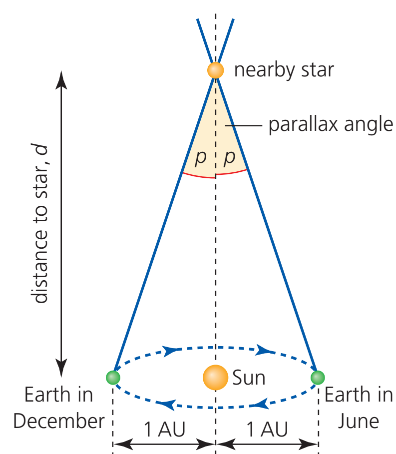
Parallax is the visual effect of a nearby object appearing to move its position, as compared to more distant objects (behind it), when viewed from different positions. A simple example is easily observed by looking at one finger held in front of your face and the background behind it, first with one eye and then the other. In the same way, a 'nearby' star can appear to *very slightly* change its position during the year compared to other stars much further away (although, as we have noted before, stars generally appear to remain in fixed patterns over very long periods of times).

Stellar parallax (Figure 16.17) was first confirmed in 1838. Many astronomers had tried to detect it before (without success) because the existence of stellar parallax provides evidence for the motion of the Earth around the Sun.

Using telescopes, astronomers measure the **parallax angle**, ***p***, between, for example, observations of the star made in December and June. Figure 16.18 shows the angular positions of a nearby star in December and June. (In Figures 16.18 and 16.19 the size of the parallax angle has been *much* exaggerated for the sake of clarity.)

■ **Figure 16.18** Measuring the parallax angle six months apart





■ **Figure 16.19** The geometry of the parallax angle

If the measurements are made exactly six months apart, the distance between the locations where the two measurements are taken is the diameter of the Earth's orbit around the Sun. We may assume that the orbit is circular, so that the radius is constant.

The parallax of even the closest stars is very small because of the long distances involved and this means that the parallax angles are so tiny that they are measured in **arc-seconds**. (There are 3600 arc-seconds in a degree.)

Once the parallax angle has been measured, simple geometry can be used to calculate the distance to the star (Figure 16.19):

$$\text{parallax angle, } p \text{ (rad)} = \frac{1.50 \times 10^{11}}{d} \text{ (m)}$$

Note that the distance from the Earth to the star and the distance from the Sun to the star can be considered to be equal for such very small angles, so: p (rad) = $\sin p = \tan p$.

Worked example

1 Calculate the distance, d, to a star if its parallax angle is 0.240 arc-seconds.

$$0.240 \text{ arc-seconds} = \left(\frac{0.240}{3600}\right) \times \left(\frac{2\pi}{360}\right) = 1.16 \times 10^{-6} \text{ rad}$$

$$p \text{ (rad)} = \frac{1.50 \times 10^{11}}{\text{d (m)}}$$

$$1.16 \times 10^{-6} = \frac{1.50 \times 10^{11}}{\text{d}}$$

$d = 1.29 \times 10^{17}$ m (= 13.7 ly)

If a parallax angle can be measured for a nearby star, calculations like this can be used to determine its distance away. Such calculations are common and it is much easier to use the angle directly as a measure of distance rather than making calculations in SI units.

We have already seen that the parsec is defined as the distance to a star that has a parallax angle of one arc-second. But there is an inverse relationship here – larger parallax angles mean smaller distances. So:

$$d \text{ (parsec)} = \frac{1}{p \text{ (arc-second)}}$$

This equation is given in the *Physics data booklet*. For example, a star with a parallax angle, p, of 0.25 arc-seconds will be from a star which is $\frac{1}{0.25} = 4$ pc away etc. Table 16.3 shows the relationship between parallax angle and distance.

The stellar parallax method is limited by the inability of telescopes on Earth to observe very small shifts in apparent positions of stars or accurately measure very small angles less than 0.01 arc-seconds. This means that this method is usually limited to those stars that are relatively close to Earth, within about 100 pc, well within our own galaxy. The use of telescopes on satellites above the turbulence and distortions of the Earth's atmosphere can extend the range considerably, but it is still not suitable for the majority of stars, which are much further away.

Table 16.3 Parallax angles in arc-seconds and distances in parsecs

Parallax angle /arc-seconds	Distance away/pc
0.10	10.00
0.25	4.00
0.50	2.00
1.00	1.00

15 Convert an angle of 1 arc-second to:
a degrees
b radians.

16 The parallax angle for Barnard's star is measured to be 0.55 arc-seconds. How far away is it from Earth:
a in pc
b in m
c in ly?

17 What are the parallax angles for three stars at the following distances from Earth?
a 2.47×10^{15} km
b 7.9 ly
c 2.67 pc

Luminosity and apparent brightness

Every star (apart from our Sun) appears to us as a point in space. The only *direct* information that we can have about any particular star is its position (as might be displayed on a *two-dimensional* star map), the intensity of radiation received from it and the spectrum of its radiation. These are the only observable differences between all the stars that we can detect.

The **apparent brightness**, ***b***, of a star (including the Sun) is defined as the *intensity* (power/area) received (perpendicular to direction of propagation) at the Earth. The units are $W\,m^{-2}$.

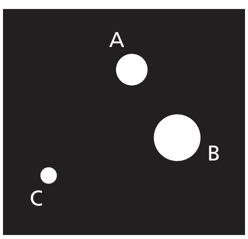


Figure 16.20 The apparent brightnesses of three stars (as indicated by the diameters of the dots)

The apparent brightness of the Sun is approximately $1360\,W\,m^{-2}$ above the Earth's atmosphere. This is also called the **solar constant** and was discussed in Chapter 8. Of course, the apparent brightnesses of all the other stars are much, much less. A typical value would be $10^{-12}\,W\,m^{-2}$. Astronomers have developed very accurate means of measuring apparent brightnesses using charge-coupled devices (CCDs), in which the charge produced in a semiconductor is proportional to the number of photons received, and hence the apparent brightness.

In Figure 16.20, stars A and B appear to be close together but in reality, in three-dimensional space, star A could be much closer to star C than star B. The situation may be further confused by differences in the brightness of the three stars. For example, it is feasible that star B could be the furthest away of these three stars and only appears brightest because it emits much more light than the other two.

The **luminosity**, ***L***, of a star is defined as the total power it radiates (in the form of electromagnetic waves). It is measured in watts, W.

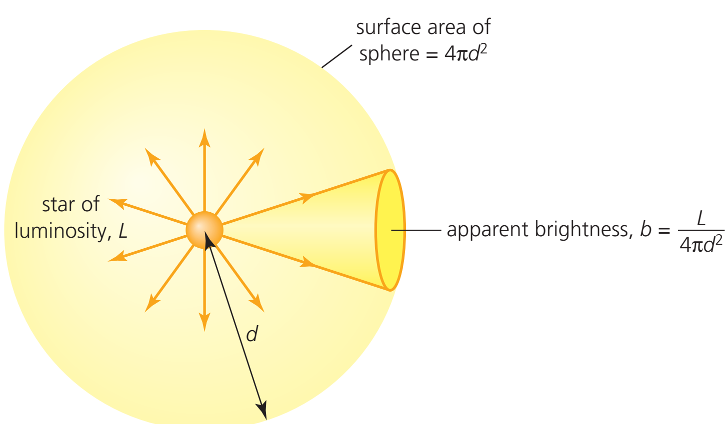


Figure 16.21 Relating apparent brightness to luminosity

For example, the luminosity of the Sun is 3.8×10^{26} W.

The apparent brightness of a star as observed on Earth will depend on its luminosity *and* its distance from Earth.

We would reasonably expect that the energy from any star spreads out equally in all directions, so the power arriving at a distant observer on Earth will be very considerably less than the power emitted. Assuming that none of the emitted energy is absorbed or scattered as it travels across space, the power received per square metre anywhere on a sphere of radius d will be equal to the emitted power (luminosity) divided by the 'surface' area of the sphere, as shown in Figure 16.21.

$$\text{apparent brightness, } b = \frac{L}{4\pi d^2}$$

This important equation is given in the *Physics data booklet*.

Nature of Science

The inverse square relationship

The importance of this equation for apparent brightness lies in the fact that once we have measured the apparent brightness of a star and *if* we know its distance from Earth, then it is a simple matter to calculate the luminosity of the star. Conversely, as we shall see later, if the luminosity of a star is known, measurement of its apparent brightness can lead to an estimate of its distance from Earth. The information provided by this simple equation is fundamental to an understanding of basic astronomy.

This is example of an **inverse square relationship**. If the distance from a star is multiplied by 2, then the apparent brightness is divided by 2^2; if the distance is multiplied by, for example, 37, then the apparent brightness will be divided by 37^2 (= 1369) etc. This is illustrated in Figure 16.22, which shows that at three times the distance, the same power is spread over nine (3^2) times the area.

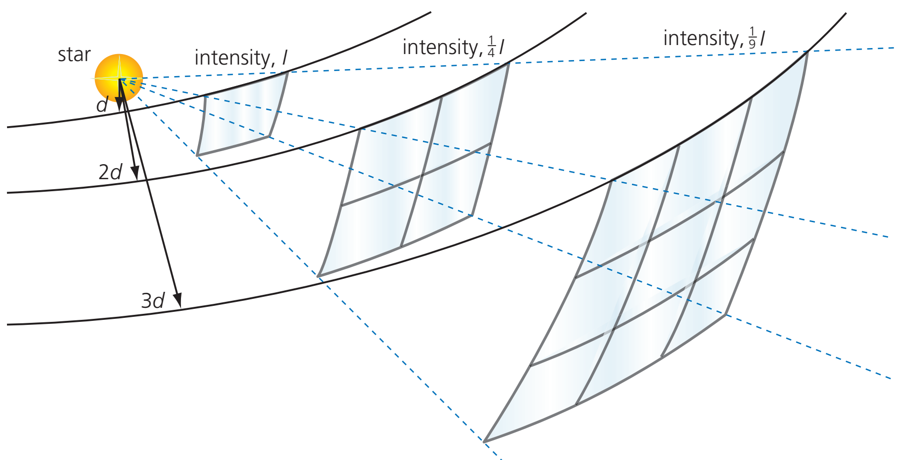


Figure 16.22 How intensity changes with the inverse square law

Not surprisingly, very little radiation is absorbed or scattered as it travels billions of kilometres through almost empty space, although the effects of the journey must be considered when studying the most distant galaxies. However, 100 km of the Earth's atmosphere does have a very significant effect, reducing brightness and resolution in many parts of the spectrum. That is why astronomers often prefer to use telescopes sited on mountain tops or on satellites above the Earth's atmosphere to gather data.

Utilizations

Telescopes on the ground and telescopes in orbit

Waves from all parts of the electromagnetic spectrum arrive at the Earth from outer space and it is truly impressive to consider just how much scientists have learned about the universe from studying these various radiations. Most of this option is about how that information is interpreted, but little has been included about how waves from the various parts of the electromagnetic spectrum provide different information about their sources. Figure 16.23 shows a telescope designed to focus and detect radio waves from outer space.

When radiation passes through the Earth's atmosphere some of it may be absorbed, refracted or scattered, and these effects will often depend on the wavelengths involved. For example,

■ **Figure 16.23** A telescope at the Very Large Array, New Mexico, USA receiving radio waves from space

■ **Figure 16.24** The telescopes at the Paranal Observatory on the top of Cerro Paranal, a mountain in the Atacama Desert in Chile

■ **Figure 16.25** The Hubble telescope

in visible light, the blue end of the spectrum is scattered more than red light and that helps to explain blue skies and red sunsets. We only have to look through the shifting haze above a hot surface to appreciate just how much convection currents in the air affect what we see.

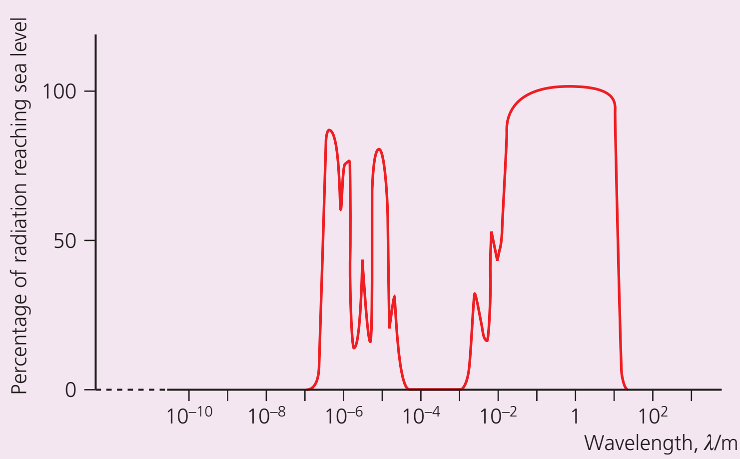
Astronomers have long understood the advantages of siting optical telescopes on the tops of mountains to reduce the adverse effects of the atmosphere on the images seen (Figure 16.24). The highest mountains are, of course, much lower than the height of the atmosphere, which is usually assumed to be approximately 100 km, although there is no distinct 'edge'.

The use of telescopes on orbiting satellites has greatly increased the *resolution* of images from space. (The resolution of images was discussed in detail in Chapter 9 and is needed by Higher Level students only.) The Hubble telescope (Figure 16.25) has been the focus of much attention, with many of its spectacular images well known around the world. The telescope was launched in 1990 and was named after the famous American astronomer, Edwin Hubble. It has a mass of about 11 tonnes and orbits approximately 560 km above the Earth's surface, taking 96 minutes for one complete orbit. One of the finest achievements of astronomers using the Hubble telescope has been determining the distances to very distant stars accurately, enabling a much improved estimate for the age of the universe.

The second major advantage of placing a satellite in orbit is that it can detect radiations that would otherwise be absorbed in the atmosphere before reaching any *terrestrial* telescopes (those on the Earth's surface). Figure 16.26 indicates (approximately) the effect that the Earth's atmosphere has on preventing radiations of different wavelengths from reaching the Earth's surface.

1 Make a sketch of Figure 16.26 and indicate and name the different sections of the electromagnetic spectrum.

2 Visit the Hubble website to look at the magnificent images from space and make a list of the important characteristics of the telescope.



■ **Figure 16.26** How the Earth's atmosphere affects incoming radiation

Worked example

2 A star of luminosity 6.3×10^{27} W is 7.9×10^{13} km from Earth. What is its apparent brightness?

$$b = \frac{L}{4\pi d^2}$$

$$b = \frac{6.3 \times 10^{27}}{4\pi \times (7.9 \times 10^{16})^2}$$

$$b = 8.0 \times 10^{-8}\,\text{W m}^{-2}$$

18 How far away from Earth is a star that has a luminosity of 2.1×10^{28} W and an apparent brightness of 1.4×10^{-8} W m^{-2}?

19 A star that is 12.4 ly from Earth has an apparent brightness of 2.2×10^{-8} W m^{-2}. What is its luminosity?

20 Calculate the distance to the Sun using values for its luminosity and apparent brightness.

21 Star A is 14 ly away from Earth and star B is 70 ly away. If the apparent brightness of A is 3200 times higher than that of star B, calculate the ratio of their luminosities.

22 If the radiation from the star in Question 18 has an average visible wavelength of 5.5×10^{-7} m, estimate how many visible photons arrive every second at a human eye of pupil diameter 0.50 cm.

Relating the luminosity of a star to its surface temperature

We know from Chapter 8 that the power radiated from a surface in the form of electromagnetic waves can be calculated from $P = e\sigma AT^4$ where A is the surface area, T is the temperature (K) and σ is the Stefan Boltzmann constant. We may assume that stars behave as *perfect black bodies*, so that emissivity $e = 1$ and the power emitted by a star (its luminosity, L) is then given by:

$$L = \sigma AT^4$$

This equation is given in the *Physics data booklet*.

Remember that when a surface is described as a 'perfect black body' we mean that it emits the maximum possible radiation at any particular temperature and not that it appears black. This equation shows us that if we know the luminosity of a star and its surface temperature then we can calculate its surface area and radius ($A = 4\pi r^2$). This is shown in the following worked example and questions. In the next section we will review how Wien's displacement law (Chapter 8) can be used to determine the surface temperature of a star from its spectrum.

Worked example

3 What is the luminosity of a star of radius 2.70×10^6 km and surface temperature 7120 K?

$$L = \sigma AT^4$$

$$= (5.67 \times 10^{-8}) \times 4\pi \times (2.70 \times 10^6 \times 10^3)^2 \times (7120)^4$$

$$= 1.33 \times 10^{28}\,\text{W}$$

23 A star has a surface area of 1.8×10^{19} m^2 and a surface temperature of 4200 K. What is its luminosity?

24 If a star has a luminosity of 2.4×10^{28} W and a surface temperature of 8500 K, what is:
- **a** its surface area
- **b** its radius?

25 What is the surface temperature of a star that has an area of 6.0×10^{20} m^2 and a luminosity of 3.6×10^{30} W?

26 If the star in question 23 is 17.3 ly away, what will its apparent brightness be when seen from Earth?

27 If the star in question 24 has an apparent brightness of 2.5×10^{-8} W m^{-2}, how many kilometres is it from Earth?

28 Compare the luminosities of these two stars: star A has a surface temperature half that of star B, but its radius is forty times larger.

29 A star has eighty times the luminosity of our Sun and its surface temperature is twice that of the Sun. How much bigger is the star than our Sun?

16.2 (D2: Core) Stellar characteristics and stellar evolution

– a simple diagram that plots the luminosity versus the surface temperature of stars reveals unusually detailed patterns that help understand the inner workings of stars; stars follow well-defined patterns from the moment they are created to their eventual death

Stellar spectra

Nature of Science

Evidence provided by spectra

Astronomers have learnt a great deal about the universe from a limited range of evidence received from sources that are enormous distances from Earth. Apart from the location and luminosity of stars, a surprisingly large amount of information can be determined from close examination of the spectrum produced by a star.

- If we can measure the wavelength at which the emitted radiation has its maximum intensity, we can calculate the *surface temperature* of a star.
- If we can observe the absorption spectrum produced by the outer layers of a star, we can determine its *chemical composition*.
- If we compare the absorption spectrum received from a star to the spectrum from the same element observed on Earth, we can use the Doppler shift to determine the *velocity* of the star (or galaxy); this provides evidence for the expansion of the universe. (See section 16.3.)

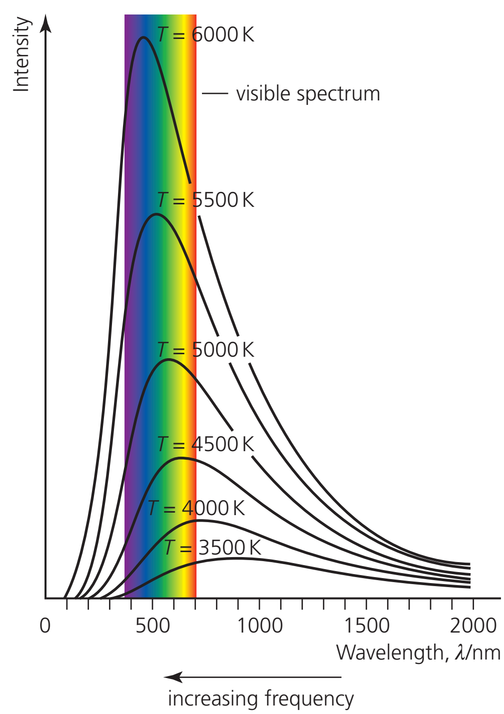


Figure 16.27 The black-body spectra emitted by stars with different surface temperatures

Surface temperature

Figure 16.27 shows that the spectra from stars with different surfaces temperatures differ, not only in overall intensity, but also in the spread of wavelengths emitted. This graph is similar to one previously seen in Chapter 8.

Wien's displacement law was also discussed in Chapter 8. It is an empirical law that represents how the wavelength at which the radiation intensity is highest becomes lower as the surface gets hotter:

$$\lambda_{max}T = 2.9 \times 10^{-3}\,\text{m K}$$

This equation is given in the *Physics data booklet*. It was also given in Chapter 8 in a slightly different form.

Worked example

4 What is the surface temperature of a star that emits radiation with a peak of intensity at a wavelength of 1.04×10^{-7} m?

$$\lambda_{max}T = 2.9 \times 10^{-3}\,\text{m K}$$

$$(1.04 \times 10^{-7})T = 2.90 \times 10^{-3}$$

$$T = \frac{2.90 \times 10^{-3}}{1.04 \times 10^{-7}}$$

$$= 27\,900\,\text{K}$$

30 If the surface temperature of the Sun is 5700 K, at what wavelength is the emitted radiation maximized? In what part of the visible spectrum is this wavelength?

31 A star emits radiation that has its maximum intensity at a wavelength of 6.5×10^{-7} m.
- **a** What is its surface temperature?
- **b** If it has a luminosity of 3.7×10^{29} W, what is the surface area of the star?
- **c** What is its radius?

32 **a** At what wavelength does a star with a surface temperature of 8200 K emit radiation with maximum intensity?
- **b** If this star has a radius of 1.8×10^{6} km, what is its luminosity?
- **c** If it is 36 ly from Earth, what is its apparent brightness?

33 The star Canopus has a luminosity of 5.8×10^{30} W and a radius of 4.5×10^{10} m. Use this data to estimate the wavelength at which it emits the most radiation.

34 Sketch graphs comparing the emission spectra from the stars Betelgeuse (3600 K) and Alkaid (20 000 K).

Utilizations

The classification of stars by the colours they emit

The surface temperatures of different stars may be as cool as a few thousand kelvin or as hot as 40 000+ K. Although Figure 16.27 only shows graphs for cooler stars, it should be clear that the range of visible colours present in the spectra from stars at different temperatures will be slightly different. For example, the light produced by a surface temperature of 4500 K has its highest intensity in the red end of the spectrum, whereas the light produced by 6000 K has more from the blue-violet end of the spectrum. Hotter stars are blue/white and cooler stars are yellow/red.

To observers on Earth this will be noticed as slight differences in colour and this has long been the way in which astronomers group and classify different stars. In general, cooler stars are slightly redder and hotter stars are slightly bluer.

Table 16.4 lists the eight spectral classes into which all visible stars are placed.

Table 16.4 Spectral classes, temperatures and colours

Spectral class	Surface temperature/K	Colour
O	30 000–50 000	blue
B	10 000–30 000	blue–white
A	7500–10 000	white
F	6000–7500	yellow–white
G	5000–6000	yellow
K	3500–5000	yellow–red (orange)
M	2000–3500	red

This apparently haphazard system of lettering stars according to their colour is an adaptation of an earlier alphabetical classification. A widely quoted mnemonic for remembering the order (from the hottest) is ‘Only Bad Astronomers Forget Generally Known Mnemonics’.

1 **a** What is the spectral class of our Sun?

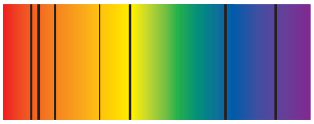
b We often refer to the light from our Sun as ‘white’. Discuss whether or not this is an accurate description.

2 Two common types of star are called red giants and white dwarfs. What spectral class would you expect them to be?

3 What is the spectral class and colour of the star Alkaid (referred to in question 34)?

Chemical composition

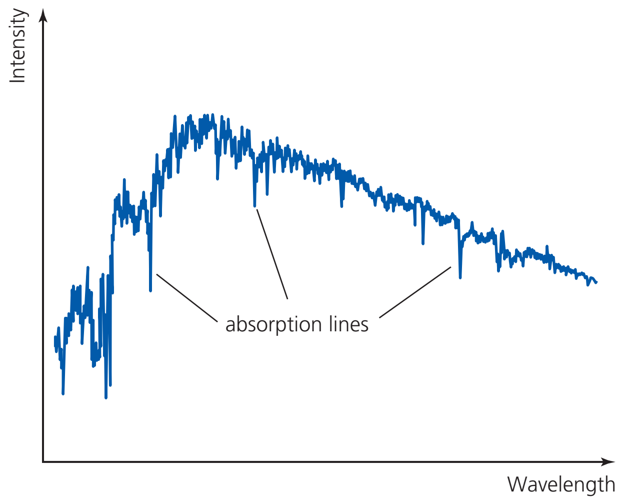
As the continuous black body spectrum emitted from a star passes through its cooler outer layers, some wavelengths will be absorbed by the atoms present. When the radiation is detected on Earth, an **absorption spectrum** (discussed in Chapter 7) will be observed.



■ **Figure 16.28** The absorption spectrum of helium

Because we know that every chemical element has its own unique spectrum, this information can be used to identify the elements present in the outer layers of a star. The element helium is the second most common in the universe (after hydrogen), but it was not detected on Earth until 1882. Fourteen years earlier, however, it had been identified as a new element in the Sun from its spectrum (see Figure 16.28).

Figure 16.29 shows a graphical representation of how a black-body spectrum emitted by the core of a star is modified by absorption of radiation in the outer layers.



■ **Figure 16.29** Graph of intensity against wavelength for a stellar absorption spectrum

ToK Link

The role of interpretation

The information revealed through spectra needs a trained mind to interpret it. What is the role of interpretation in gaining knowledge in the natural sciences? How does this differ from the role of interpretation in the other areas of knowledge?

Without detailed scientific knowledge and understanding, the observation of spectra would offer no obvious clues about the nature of stars. This is equally true of many other aspects of astronomy. Without scientific expertise, the information is of no use and may seem irrelevant, so that a non-expert would probably be unable to comment meaningfully. The same comments apply to advanced studies in other scientific disciplines.

■ The Hertzsprung–Russell (HR) diagram

The luminosity of a star depends on its surface temperature and its surface area ($L = \sigma AT^4$), so a star could be particularly luminous because it is hot or because it is big, or both.

In the early twentieth century two scientists, Hertzsprung and Russell, separately plotted similar diagrams of luminosity against temperature in order to determine whether or not there was any pattern in the way that the stars were distributed.

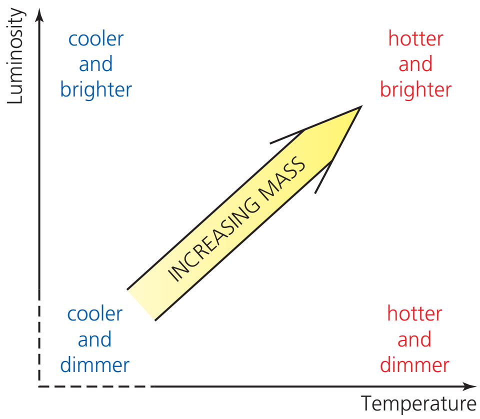
If there were no similarities in the composition of different stars, they could have many different combinations of temperature, size and luminosity. This would lead to them being randomly distributed on a luminosity–temperature diagram. But, more than about 90% of all stars are undergoing the same processes, fusing hydrogen into helium (as explained previously) and they are in a similar kind of equilibrium. These stars are called **main sequence stars** and their only essential difference is their mass.

Mass–luminosity relation for main sequence stars

Stars that are formed from *higher masses* will have stronger gravitational forces pulling them together. This will result in higher temperatures at their core and faster rates of nuclear fusion. More massive main sequence stars will have larger sizes, higher surface temperatures and *brighter luminosities*.

The relationship between luminosity, L, and mass, M, for main sequence stars is described by the following equation, which is given in the *Physics data booklet*. This is an approximate, generalized relationship and it may not be precise for any given star.

$$L \propto M^{3.5}$$



■ **Figure 16.30** Linking mass, temperature and luminosity for main sequence stars

For example, if star A has twice the mass of star B, star A will have a luminosity approximately $2^{3.5}$ times greater than star B (≈ ×11). This means that the rate of nuclear fusion in the more massive star will be much higher and it will have a much shorter lifetime than a less massive star.

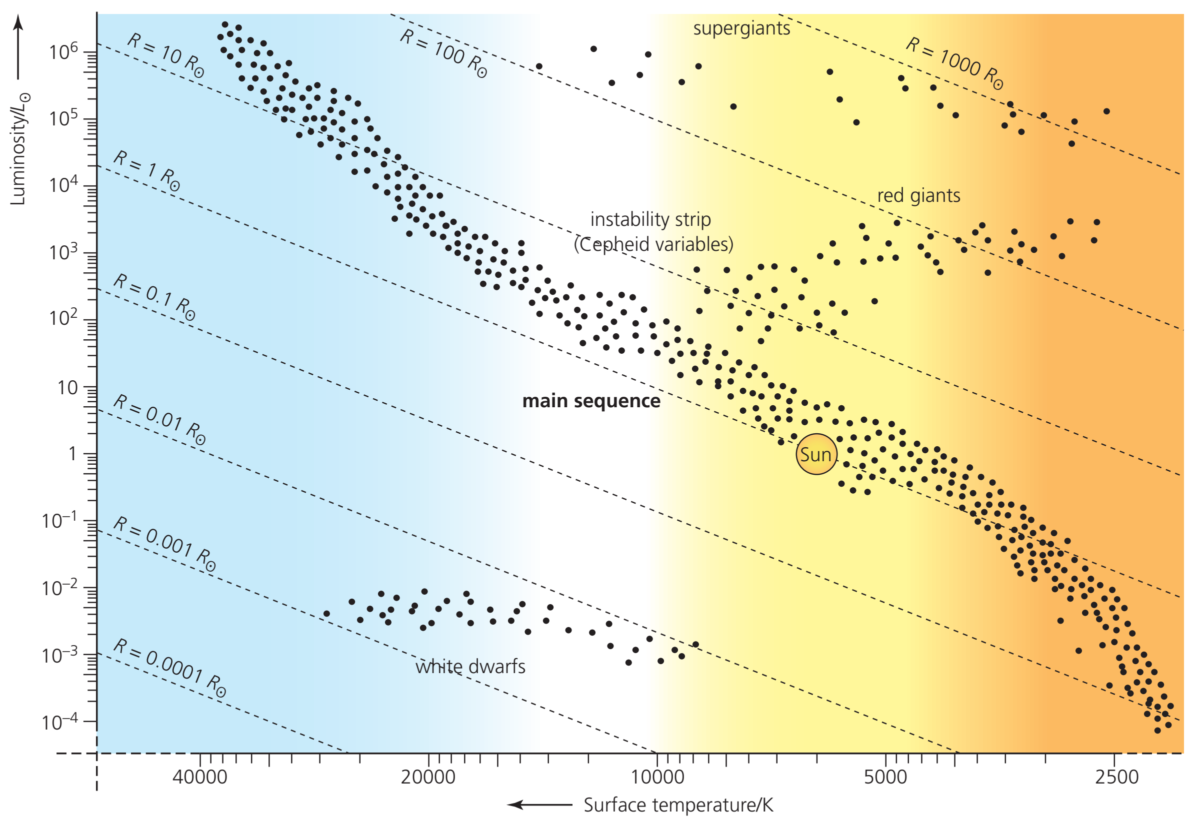
If the relationship between the mass and luminosity of a star is represented by this relationship, then we can be sure that it is a main sequence star.

Figure 16.30 suggests how we might expect a luminosity–temperature diagram to appear for main sequence stars of different masses.

Hertzsprung and Russell plotted data from a very large number of stars on luminosity–temperature diagrams, but the important HR diagram has two significant differences from Figure 16.30:

1 For historical reasons the temperature scale is reversed.
2 Because of the enormous differences in the luminosity of stars, the scale is logarithmic, rather than linear. (The temperature scale is also usually logarithmic.)

Figure 16.31 shows a large number of individual stars plotted on a Hertzsprung–Russell (HR) diagram, with all luminosities compared to the luminosity of our Sun ($L_{\odot}$). This figure also tries to give an impression of the colours of the stars.



■ **Figure 16.31** The Hertzsprung–Russell (HR) diagram

It should be apparent that the stars are *not* distributed at random in the HR diagram. The diagram can be used as a basis for classifying stars into certain types.

As already explained, most stars (about 90%) can be located in the central band, which runs from top left to bottom right in Figure 16.31. These are the main sequence stars. The 10% of stars which are *not* in the main sequence are important and they will be discussed in the next section. In general, we can say that any stars located vertically *above* the main sequence must be larger (than main sequence stars) in order to have higher luminosity at the same temperature. By similar reasoning, any stars vertically *below* the main sequence must be smaller than main sequence stars of the same temperature.

By considering $L = \sigma AT^4$ and $A = 4\pi R^2$ (leading to $L = \sigma 4\pi R^2T^4$), it is possible to draw the lines of constant radius on the HR diagram (as shown in Figure 16.31).

Worked example

5 Use the HR diagram in Figure 16.31 to predict the surface temperature of a main sequence star that has ten times the radius of our Sun.

The band of main sequence stars crosses the $R = 10R_\odot$ line at about 30 000 K.

35 The radius of the Sun is 7×10^8 m and its surface temperature is 5800 K. Estimate the radius of a main sequence star that has a surface temperature five times that of our Sun.

36 The luminosities of two main sequence stars are in the ratio 10 : 1. What is the ratio of their masses?

37 **a** A star has a mass five times heavier than the mass of our Sun. Estimate its luminosity.
b What assumption did you make?
c Which star will have the longer lifetime?
d Use the HR diagram in Figure 16.31 to estimate the surface temperature of the star.
e Approximately how many times bigger is this star than our Sun?

38 Use the HR diagram to estimate the difference in diameter of a white dwarf star and a supergiant star if they both have the same surface temperature.

Utilizations

Using the HR diagram to estimate the distance to stars

Using Wien's law the surface temperature of a star can be determined from its spectrum and, assuming that it is a main sequence star, it is a relatively simple matter to use the HR diagram to estimate its luminosity, L, and hence calculate its distance, d, away from Earth using $b = L/4\pi d^2$ and a measurement of its apparent brightness, b.

This method assumes that the radiation, which has travelled vast distances from very distant stars, has not been altered in any way by the journey. For example, if any radiation is absorbed or scattered during the journey, the value of apparent brightness used in calculations will be less than it would have been without the absorption or scattering, leading to an over-estimate of the distance to the star.

Because the exact position of the star on the HR diagram may not be known with accuracy and because of unknown amounts of scattering/absorption, there is a significant uncertainty in this method of determining stellar distances.

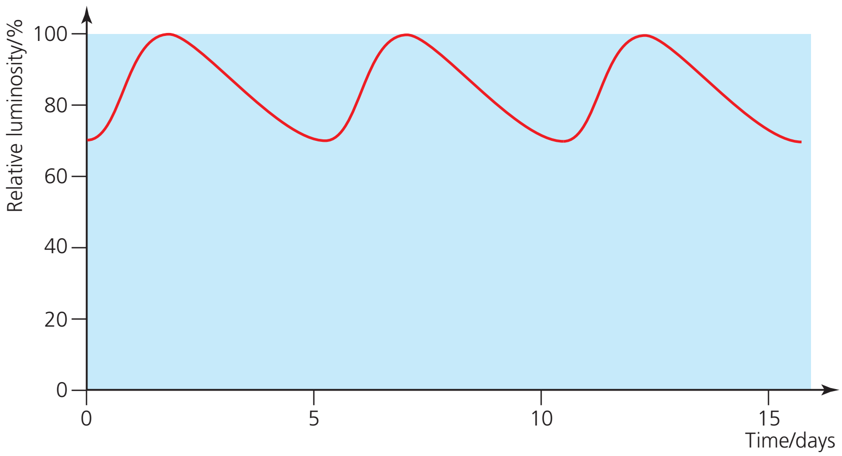
In fact, the use of this method (misleadingly called *spectroscopic parallax*) is mostly confined to our galaxy. The majority of stars are obviously further away in other galaxies, so to determine the distances to those galaxies we need other methods.

1 Estimate the distance from Earth (in pc) of a main sequence star that has a surface temperature of 7500 K and an apparent brightness of 4.6×10^{-13} W m^{-2}.

Types of star that are *not* on the main sequence

Cepheid variables

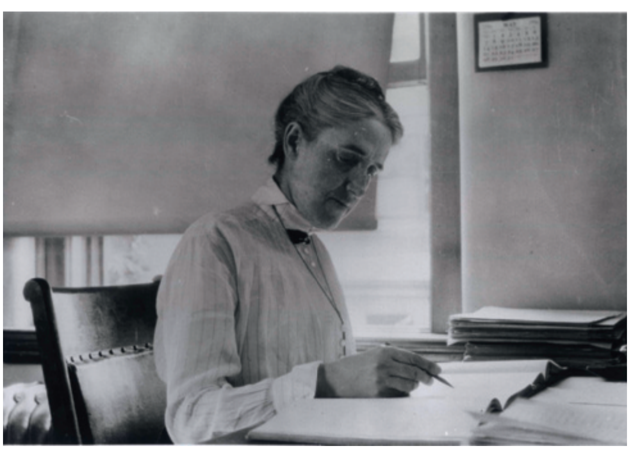
The **instability strip** on the HR diagram contains a number of different kinds of pulsating stars. Such stars have moved off the main sequence and are oscillating under the effects of the competing gravitational pressure, and radiation and thermal pressures. The most important stars in the instability strip are known as **Cepheid variables**.



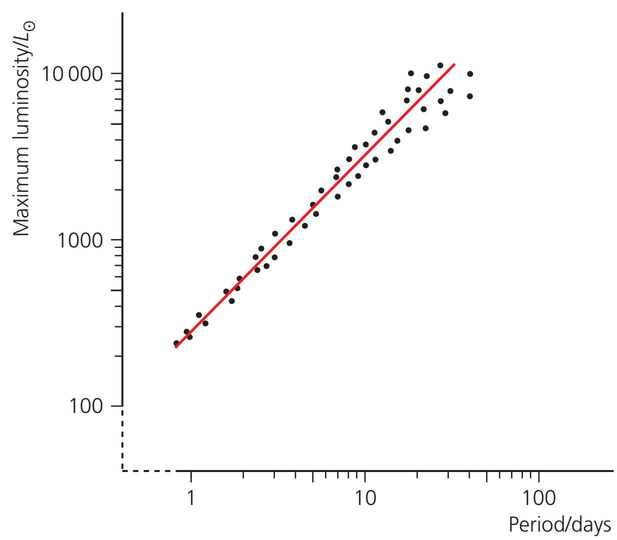
■ **Figure 16.32** Variation in luminosity of a Cepheid variable star

In a Cepheid variable the outer layers regularly expand and contract (typically by 30%) with surprisingly short periods (in astronomical terms), resulting in very regular and precise variations in luminosity (see Figure 16.32) – a typical period is a few weeks. If the surface temperature remains approximately constant then the increasing luminosity is explained by the larger surface area when the star expands.

Although Cepheid variables are not common stars, they are important and their behaviour has been studied in great depth. From observations of those Cepheid variables that are relatively close to Earth, it is known that there is a precise relationship between the time period of their pulses of luminosity (and hence their received apparent brightness on Earth) and the peak value of that luminosity. This was first discovered by Henrietta Leavitt (Figure 16.33) in 1908. This relationship is called the **period–luminosity relationship** and it is commonly presented in graphical form, as shown in Figure 16.34; the longer the period, the higher the luminosity of the Cepheid variable.



■ **Figure 16.33** Henrietta Leavitt discovered the periodicity of Cepheid variables in 1908.



■ **Figure 16.34** Period–luminosity relationship for a Cepheid variable

Note the logarithmic nature of both the scales on the graph in Figure 16.34. This is necessary to include the enormous range of values involved.

Using Cepheid variables to determine astronomical distances

If the luminosity of a Cepheid variable can be determined from its period, then its distance from Earth, d, can be calculated if its apparent brightness, b, is measured. Again we can use the equation:

$$b = \frac{L}{4\pi d^2}$$

Inaccuracies in the data involved mean that these estimates of distance, especially to the furthest galaxies, are uncertain. This uncertainty is a significant problem when estimating the age of the universe.

Astronomers often describe Cepheid variables as '**standard candles**' because if their distance from Earth is determined, it can then be taken as a good indication of the distance of the whole galaxy from Earth (since that distance is very much longer than the distances between stars within the galaxy, see Figure 16.15).

Worked example

6 A Cepheid variable in a distant galaxy is observed to vary in apparent brightness with a period of 8.0 days. If its maximum apparent brightness is $1.92 \times 10^{-9}\,W\,m^{-2}$, how far away is the galaxy?

From a luminosity–period graph (similar to Figure 16.34), the maximum luminosity can be determined to be 2500 times the luminosity of the Sun.

luminosity = $2500 \times (3.8 \times 10^{26}\,W) = 9.5 \times 10^{29}\,W$

$$b = \frac{L}{4\pi d^2}$$

$$1.92 \times 10^{-9} = \frac{9.5 \times 10^{29}}{4\pi d^2}$$

$d = 6.3 \times 10^{18}\,m$

39 **a** If a Cepheid variable has a period of 15 days, what is its approximate maximum luminosity?
b If the star is 3.3 Mpc from Earth, what is the maximum observed apparent brightness?

40 A Cepheid variable is 15 kpc from Earth and is observed to have a maximum apparent brightness of $8.7 \times 10^{-13}\,W\,m^{-2}$.
a Calculate the maximum luminosity of this star.
b Use Figure 16.34 to estimate the time period of the variation in the star's luminosity.

41 For very large distances astronomers may use supernovae (rather than Cepheids) as 'standard candles'. Suggest a property of supernovae which might be necessary for this.

What happens to a star when the supply of hydrogen is reduced?

Over a long period of time, the amount of hydrogen in the core of a star gets significantly less, so that eventually the outwards pressure is reduced and becomes smaller than the inwards gravitational pressure. This occurs when the mass of the core is about 12% of the star's total mass and there is still plenty of hydrogen remaining in the outer layers of the star. The star begins to contract and gravitational energy is again transferred to kinetic energy of the particles (the temperature of the core rises even higher than before – to 10^8 K and higher). This forces the outer layers of the star to expand and consequently cool.

At the higher temperatures in the core (in all but the very smallest stars) it is now possible for helium to fuse together to form carbon and possibly some larger nuclei, releasing more energy so that the star becomes more luminous. So, the star now has a hotter core but it has become larger and cooler on the surface. Its colour changes and it is now known as a **red giant** (or, if it is very large, a **red supergiant**). At this point it will leave the main sequence part of the HR diagram. All main sequence stars follow predictable patterns but the heavier the mass of a star, the higher the gravitational potential energy and the higher the temperature when it begins to collapse. We can identify three different outcomes, depending only on the mass of the original star: white dwarf, neutron star or black hole.

Red giants, white dwarfs, neutron stars and black holes

Red giants

As explained above, most stars will become red giants (or red supergiants) at the end of their time on the main sequence. They are described as *giant* stars because they have expanded considerably from their original size and, in doing so, their surfaces have cooled and therefore changed in colour to slightly red.

White dwarfs

After nuclear fusion in the core finishes, if the mass of a red giant star is less than a certain limit (about eight solar masses), the energy released as the core contracts forces the outer layers of the star to be ejected in what is known as a **planetary nebula**. (Be careful with this name – it is misleading because it has nothing to do with planets.) The core of the star that is left behind has a much reduced mass and is described as a *dwarf star*.

A process known as **electron degeneracy pressure** (electrons acting like a gas) prevents the star collapsing further so this kind of star can remain stable for a long time. Such stars are known as **white dwarfs** because they have low luminosities (they cannot be seen without a telescope), but their surface temperatures are relatively high ($L = \sigma AT^4$).

Studying the patterns we see in other stars helps us to understand our own Sun and what will happen to it in the future. It is about halfway through its time as a main sequence star and it will become a red giant in about seven billion years, after which it will become a white dwarf.

Neutron stars or black holes

Red giants with original masses greater than about eight solar masses are known as red supergiants and they do *not* evolve into white dwarfs. The electron degeneracy pressure is insufficient to resist the gravitational forces and the gravitational potential energy released is so high that there are dramatic changes in the core that result in an enormous explosion called a **supernova**. Here again, the result depends on the mass involved. If the original mass of the star was between 8 and 20 solar masses, the remaining core after the supernova will form a **neutron star**. If the mass was greater, a **black hole** is formed.

Neutron stars

After a supernova of a red supergiant, if the remaining core has a mass between approximately 1.4 and 3 solar masses it will contract to a neutron star. Neutron stars are extremely dense ($\rho \approx 5 \times 10^{17}\,\mathrm{kg\,m^{-3}}$), but resist further compression due to a process called **neutron degeneracy pressure**.

Black holes

If the remnant after a supernova has a remaining mass of more than approximately three solar masses, neutron degeneracy pressure is insufficient to resist further collapse. The result is a black hole, which produces such strong gravitational forces that not even the fastest particles, photons (for example, light) can escape.

Black holes cannot be observed directly, but they can be detected by their interaction with other matter and radiation. For example X-rays are produced when superheated matter spirals towards a black hole. NASA's Chandra Observatory was designed to search for black holes.

The first black hole was confirmed in 1971. Astronomers believe that our own galaxy, the Milky Way, has a supermassive black hole near its centre.

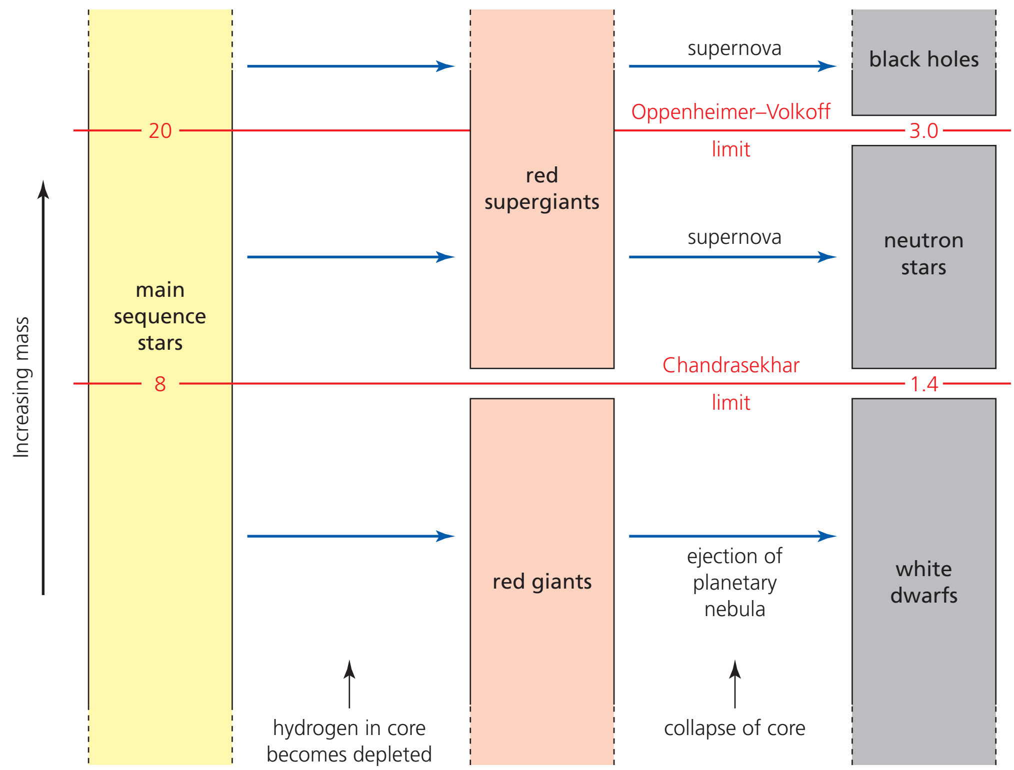
Chandrasekhar and Oppenheimer–Volkoff limits

The mass limits mentioned above are known by the names of prominent astronomers:

- The Chandrasekhar limit is the maximum mass of a white dwarf star (= 1.4 × solar mass).
- The Oppenheimer–Volkoff limit is the maximum mass of a neutron star (≈ 3 × solar mass).

Figure 16.35 represents these limits in a simplified chart of stellar evolution.

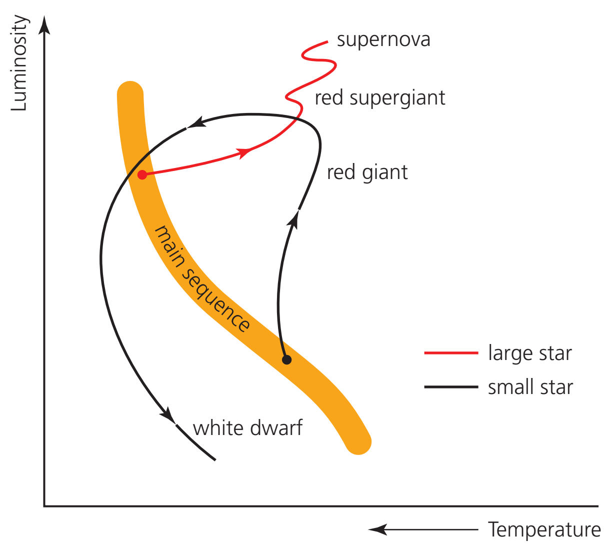
■ **Figure 16.35** Evolution of stars of different masses (the numbers shown represent the approximate mass limits of the stars as multiples of the current mass of the Sun)



■ Stellar evolution on HR diagrams

When a main sequence star expands to a red giant, or a red supergiant, its luminosity and surface temperature change and this, and subsequent changes over time, can be tracked on an HR diagram. It is known as a star's **evolutionary path**. Typical evolutionary paths of low-mass and high-mass stars are shown in Figure 16.36.

■ **Figure 16.36** Evolutionary paths of stars after they leave the main sequence



42 Explain why neutron stars and black holes cannot be placed on a HR diagram.

43 Use the internet to learn more about electron and neutron degeneracy.

44 Suggest how a black hole can be detected, even though it cannot be seen.

45 Explain why the Chandrasekhar limit is such an important number in astronomy.

46 Explain why some supernovae result in neutron stars, while others result in black holes.

16.3 (D3: Core) Cosmology *– the Hot Big Bang model is a theory that describes the origin and expansion of the universe and is supported by extensive experimental evidence*

Cosmology is the study of the universe – how it began, how it developed and what will happen to it in the future. It has always been the nature of many individuals, societies and cultures to wonder what lies beyond the Earth. The fact that the Sun and the stars appear to revolve around the Earth led early civilizations, understandably but wrongly, to believe that a stationary Earth was the centre of everything. This belief was often fundamental to their religions. Indeed, even today some people still believe from their everyday observations, or their religious beliefs, that the Sun orbits around the Earth rather than the other way around.

Nature of Science

Models of the universe

In the **Newtonian model of the universe**, the Earth, the Sun and the planets were just tiny parts of an infinitely large and unchanging (static) universe that had always been the way it is, and always would be the same. In this model, the universe, on the large scale, is more or less the same everywhere. In other words it is uniform with stars approximately evenly distributed. Newton reasoned that unless all of these assumptions (sometimes called postulates) were valid, then gravitational forces would be unbalanced, resulting in the movement of stars (which were thought to be stationary at that time).

But there is a big problem with this Newtonian model of the universe, one that many astronomers soon realized. If the universe is infinite and contains an infinite number of stars, there should be no such thing as a dark sky at night, because light from the stars should be arriving from all directions at all times. (This is known as *Olbers's paradox*, named after one of the leading astronomers of the nineteenth century, Heinrich Wilhelm Olbers. A paradox is an apparently true statement that seems to contradict itself. 'I always lie' is a widely quoted paradoxical statement.)

It was clear that either the reasoning given above and/or the Newtonian model of the universe needed changing or rejecting. Since the mid to late 1960s, the **Big Bang model** of the universe has been widely accepted by astronomers and has solved Olbers's paradox.

Additional Perspectives

'The shoulders of giants'

Nicolas Copernicus, a Polish astronomer and cleric (Figure 16.37), is considered by many to be the founder of modern astronomy. In 1530 he published a famous paper stating that the Sun was the centre of the universe and that the Earth, stars and planets orbited around it (a *heliocentric model*). At that time, and for many years afterwards, these views directly challenged 'scientific', philosophical and religious beliefs. It was then generally believed that the Earth was at the centre of everything (a *geocentric model*). That profound and widespread belief dated all the way back to Ptolemy, Aristotle and others nearly 2000 years earlier. It should be noted, however, that Aristrachus in Ancient Greece is generally credited with being the first well-known person to propose a heliocentric model.

■ **Figure 16.37** Copernicus

More than 100 years after the birth of Copernicus, and still before the invention of the telescope, an eccentric Danish nobleman, Tycho Brahe, became famous for the vast number of very accurate observations he made on the motions of the five visible planets. He worked mostly at an elaborate observatory on an island in his own country, but went to Prague a few years before his death in 1601.

Johannes Kepler was Brahe's assistant and later, after his death, he worked on Brahe's considerable, but unexplained, data to produce his three famous laws of planetary motion.

At about the same time in Italy, the astronomer Giordano Bruno had taken the heliocentric model further with revolutionary suggestions that the universe was infinite and that the Sun was not at the centre. The Sun was, Bruno suggested, similar in nature to the other stars. He was burned at the stake in 1600 for these beliefs – killed for his, so-called, heresy. About 30 years later, one of the greatest scientific thinkers of all time, Galileo Galilei, was placed on trial by the Roman Catholic Church under similar charges. Many years earlier he had used the newly invented telescope to observe the moons of Jupiter and had reasoned that the Earth orbited the Sun in a similar way, as had been proposed by Copernicus. Under pressure, he publicly renounced these beliefs and was allowed to live the rest of his life under house arrest. All this has provided the subject of many books, plays and movies.

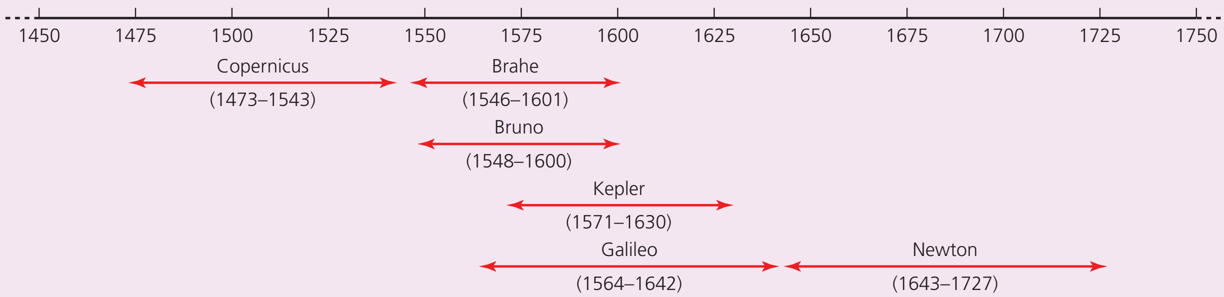


Figure 16.38 Time line of some famous early astronomers

Although Kepler had found an accurate way to describe the motion of the planets mathematically, an explanation was not produced until about 80 years later (Figure 16.38) when Newton was able to use the motion of the planets and the Moon as evidence for his newly developed theory of universal gravitation (Chapter 6).

1 Many people would place Newton and Galileo in a list of the top five scientists of all time but, to a certain extent, that is just a matter of opinion.
   a Why do you think Newton and Galileo are so highly respected?
   b What criteria would you consider when trying to decide if a scientist was 'great'?
2 Research the origins of the quotation 'the shoulders of giants', which forms the title of this Additional Perspectives section.

The Big Bang model

The Big Bang model is the current theory about how the universe began at one precise moment in time, 13.8 billion years ago. Before looking more closely at this theory, we will first consider the evidence for an **expanding universe**.

Using spectra to determine the velocity of stars and galaxies

If a source of light is not stationary but moving towards or away from an observer, there will be a *shift* (a very slight change) in all the wavelengths and frequencies of the light received. This is similar to the Doppler effect in the sound received from moving vehicles – as a police car approaches, we hear a higher-pitched sound (shorter wavelength) than when it is moving away from us (Figure 16.39).

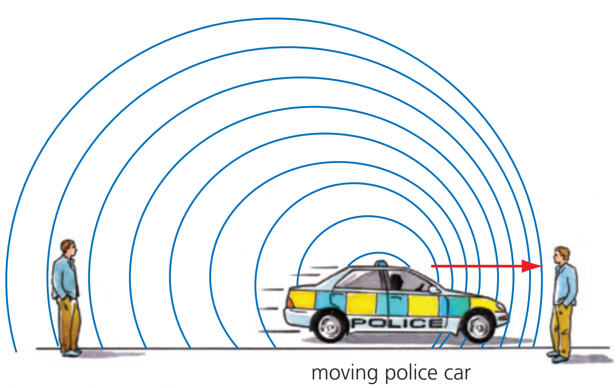


Figure 16.39 The Doppler effect for sound

In the case of light waves, the shift is very small and is usually undetectable unless a source is moving *very* quickly, such as a star or galaxy. In order to detect a shift for light we need to examine the line spectrum from the source and compare it to the line spectrum produced by the same element(s) on Earth.

We find that the *pattern* of the absorption lines on a spectrum is the same, but all the lines are very slightly shifted from the positions they would occupy if there were no motion of the source relative to the observer. Careful observation of the line spectrum received from a star (Figure 16.40) can be used to calculate the velocity of the star. In example A in Figure 16.40, all the absorption lines have been shifted towards lower frequencies and this is commonly described as a **red-shift**. A red-shift occurs in the radiation received from a star or galaxy that is moving away (*receding*) from the Earth. If a star or galaxy is moving towards Earth, then the shift will be towards higher frequencies and is called a **blue-shift**, as shown in example B. (This is unusual for galaxies.)

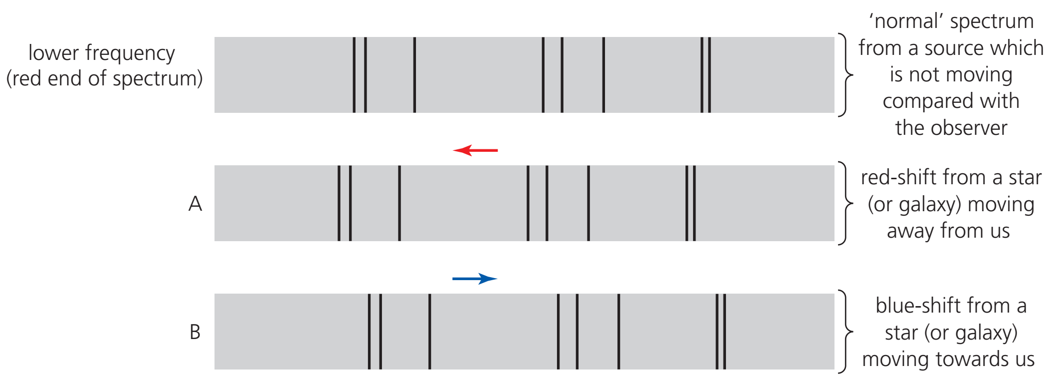


Figure 16.40 Red- and blue-shifts

For a given wavelength, λ_0, in a line spectrum, the shift (difference) in wavelength, $\Delta\lambda$, received from a fast-moving star or galaxy is proportional to its speed towards or away from the observer. The ratio of $\Delta\lambda/\lambda_0$ is the numerical representation of red-shift and is given the symbol z. For a speed v, which is significantly slower than the speed of light, c, the red-shift, z, is given by the equation

$$z = \frac{\Delta\lambda}{\lambda_0} \approx \frac{v}{c}$$

This equation is given in the *Physics data booklet* and is similar to the equation used in Chapter 9. Because it is a ratio, red-shift does not have a unit.

If we can measure the red shift for a known wavelength, we can calculate the recession speed of the source (star or galaxy). Basic calculations like these assume that the source of light is moving in a straight line directly away from the Earth. As we shall see, this *is* a reasonable assumption, although it is not necessarily perfectly true.

Worked example

7 A line in the hydrogen spectrum has a wavelength of 4.34×10^{-7} m. When detected on Earth from a distant galaxy, the same line has a wavelength of 4.76×10^{-7} m. What is the speed of the galaxy?

$\Delta\lambda = (4.76 \times 10^{-7}) - (4.34 \times 10^{-7}) = 4.2 \times 10^{-8}$ m

$$z = \frac{\Delta\lambda}{\lambda_0} = \frac{v}{c}$$

$$\frac{4.2 \times 10^{-8}}{4.34 \times 10^{-7}} = \frac{v}{3.00 \times 10^{8}}$$

$v = 2.90 \times 10^{7}\ \text{m s}^{-1}$

Because the shift is to a longer wavelength (a red-shift), we know that the motion of the galaxy is *away* from Earth. We say that the galaxy is *receding* from Earth. When the light from a large number of galaxies is studied, we find that *nearly all* the galaxies are receding from Earth and each other. This can only mean that the universe is expanding.

47 What is the red-shift of a galaxy with a recession speed of:
- **a** $2.2 \times 10^6\,\text{m s}^{-1}$
- **b** 10% of the speed of light?

48 What is the recession speed of a galaxy (km h^{-1}) if radiation of original wavelength 6.5×10^{-7} m undergoes a red-shift of 3.7×10^{-8} m?

49 A star receding at a velocity of $9.2 \times 10^3\,\text{km s}^{-1}$ emits radiation of wavelength 410 nm. What is the extent of the red-shift of this radiation when it is received on Earth and what is its received wavelength?

50 Hydrogen emits radiation of frequency 6.17×10^{16} Hz. What frequency will be detected on Earth from a galaxy moving away at $1.47 \times 10^7\,\text{m s}^{-1}$?

51 Only a very tiny percentage of galaxies are moving towards us. Research the blue-shift of the Andromeda galaxy, one of the galaxies in the Local group.

Hubble's law

In the mid-1920s, the American astronomer Edwin Hubble compared information about the **recession speeds** of relatively nearby galaxies (obtained from the red-shift of the light received) with the distances of the galaxies from Earth that were determined by using Cepheid variables within the galaxies. By 1929 Hubble had gathered enough data to publish a famous graph of his results for Cepheids within distances of a few Mpc from Earth. Figure 16.41 includes more results and for greater distances.

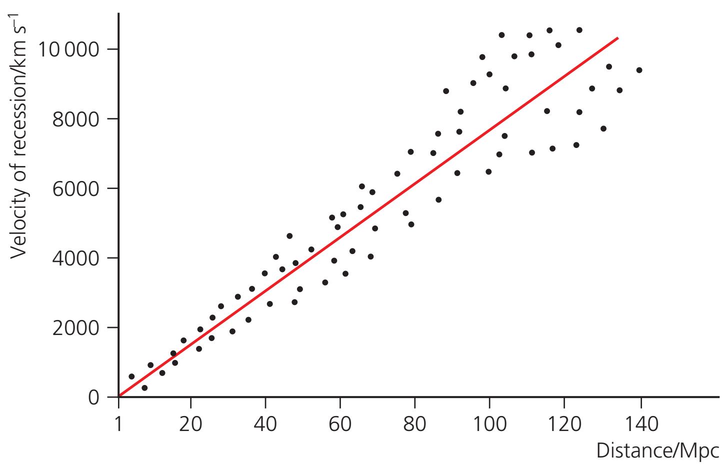


Figure 16.41 Variation of recession speeds of galaxies with their distances from Earth

Even today there are significant uncertainties in the data represented on this graph (although error bars are not shown on Figure 16.41). These uncertainties are mainly because the precise measurement of distances to galaxies is difficult, but also because galaxies move within their clusters. Nevertheless, the general trend is very clear and was first expressed in **Hubble's law**:

The current velocity of recession, v, of a galaxy is proportional to its distance, d (from Earth).

This can be written as:

$$v = H_0 d$$

This equation is listed in the *Physics data booklet*.

H_0 is the gradient of the graph and is known as the **Hubble constant**. Because of the uncertainties in the points on the graph, the Hubble constant is not known accurately, despite repeated measurements. The currently accepted value is about $70\,km\,s^{-1}\,Mpc^{-1}$ (this unit is more widely used than the SI unit, s^{-1}). However, different determinations of the Hubble constant have shown surprising variations. Hubble's 'constant' is believed to be a constant for everywhere in the universe at this time, but over the course of billions of years its value has changed.

Worked example

8 Estimate the gradient of the graph in Figure 16.41 and compare it with the value given for the Hubble constant in the previous paragraph.

$$\text{gradient, } H_0 = \frac{v}{d} = \frac{9000}{120}$$
$$= 75\,km\,s^{-1}\,Mpc^{-1}$$

This value varies by 7% from the value quoted earlier, but neither figure includes any assessment of uncertainty, so it is possible that they are consistent with each other.

Hubble's law can be applied to the radiation received from all galaxies that are moving free of significant 'local' gravitational forces from other galaxies. That is, the law can be used for isolated galaxies or clusters (considered as one object), but is less accurate for individual galaxies moving within a cluster because the resultant velocity of an individual galaxy is the combination of its velocity with respect to the cluster and the recession velocity of the cluster as a whole. A few galaxies even have a resultant velocity *towards* the Earth at this time and radiation received from such galaxies is blue-shifted.

The use of the Hubble constant with the recessional speeds of distant galaxies provides astronomers with another way of calculating the distance to galaxies which are too far away to use alternative methods.

More about the Big Bang

The conclusion from Hubble's observations can only be that the universe is expanding because (almost) all galaxies are moving away from the Earth.

It is important to realize that this is true for galaxies observed in *all* directions and *would also be true for any observer viewing galaxies from any other location in the universe*. Almost all galaxies are moving away from all other galaxies. Our position on Earth is not unique, or special, and we are not at the 'centre' of the universe – the universe does not have a centre.

Calculations confirm that the further away a galaxy is, the faster it is receding. This simple conclusion has very important implications: the more distant galaxies are further away *because* they travelled faster from a common origin. Observations suggest that all the material that now forms stars and galaxies originated at the same place at the same time. An expanding model of the universe had been proposed a few years earlier by Georges Lemaître and this was developed in the 1940s into what is now called the Big Bang model.

If radiation from a star or galaxy is observed to have a blue-shift, it is because it is moving towards Earth. This is not evidence against the Big Bang model because such an object is moving within a gravitationally bound system (a galaxy, a cluster of galaxies or a binary star system) and at the time of observation it was moving towards the Earth faster than the system as a whole was moving away. For example, our neighbouring galaxy, Andromeda, exhibits a small blue-shift – it is moving towards us as part of its motion within our local group of galaxies, which is a gravitationally bound system.

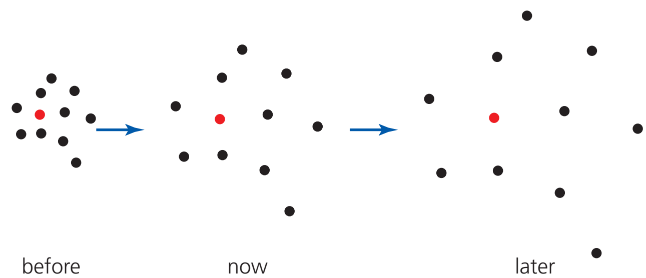
In the Big Bang model, the universe was created at a point about 13.8 billion (1.38×10^{10}) years ago. At that time it was incredibly dense and hot, and ever since it has been expanding and cooling down.

The expansion of the universe is the expansion of space itself and it should not be imagined as being similar to an explosion, with fragments flying into an existing space (void), like a bomb exploding.

It may be helpful to visualize the expansion of space using marks on a very large rubber sheet to represent galaxies. (Imagine that the sheet is so large that the edges cannot be seen.) If the sheet is stretched equally in all directions, all the marks move apart from each other. Of course, a model like this is limited to only two dimensions (Figure 16.42).

The red-shift of light should be seen as a consequence of the expansion of space rather than being due to the movement of galaxies through a fixed space.

Figure 16.42 An expanding universe



It is very tempting to ask 'what happened before the Big Bang?' In one sense, this question may have no answer because the human concept of time is all about change – and before the Big Bang there was nothing to change.

The Big Bang should be considered as the creation of *everything* in our universe – matter, space and time.

Nature of Science

Simplicity

Expressed in basic terms, the Big Bang model of the universe is elegant in its simplicity. In judging scientific theories and models, as well as other human endeavours, simplicity is often (but not always) an admirable aim. This has been expressed in what is known as **Occam's razor** – if you need to choose between two or more possible theories, select the one with the fewest assumptions. Until you know that a more complicated theory is preferable, simplicity may be the best criterion to judge between opposing models.

More complicated theories are more difficult to test and, if they are found to be in doubt, it is often possible to add another layer of (unproven) theory to retain some of their credibility.

Age of the universe

We can make an estimate of the time since the Big Bang (the age of the universe) using Hubble's constant ($70\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1}$).

Because time, t = distance, d/velocity, v and $v = H_0 d$, we can write:

$$T \approx \frac{1}{H_0}$$

where T is the approximate age of the universe, often called **Hubble time**. This equation is given in the *Physics data booklet*.

Calculated using this equation, T can be considered to be an approximate and upper limit to the age of the universe for the following reasons.

- It is not sensible to assume that the recession speed of the galaxies has always been the same. It is reasonable to assume that the speed of galaxies was fastest in the past when they were closer together and that they are now slowing down because of gravitational attraction. (We now know that this is not true: discussed in more detail later.)
- We do not know that the expansion started at the same time as the Big Bang.
- The uncertainty in the Hubble constant is significant.

We would prefer the time to be in SI units, and in SI units Hubble's constant becomes:

$$\frac{70 \times 10^3}{3.26 \times 10^6 \times 9.46 \times 10^{15}} = 2.27 \times 10^{-18}\,\text{s}$$

so that

$$T = \frac{1}{H_0} = \frac{1}{2.27 \times 10^{-18}} = 4.4 \times 10^{17}\,\text{s (or } 1.4 \times 10^{10} \text{ years).}$$

52 What is the recession speed (km s^{-1}) of a galaxy that is 75 Mpc from Earth?

53 How far away is a galaxy travelling at 1% of the speed of light?

54 Galaxy A is a distance of 76 Mpc from Earth and is receding at a velocity of $5500\,\text{km s}^{-1}$. Another galaxy, B, is receding at $7300\,\text{km s}^{-1}$. Without using a value for H_0, estimate the distance to galaxy B.

55 A spectral line of normal wavelength 3.9×10^{-7} m is shifted to 4.4×10^{-7} m when it is received from a certain distant galaxy.
a How fast is the galaxy receding?
b How far away is it?

Cosmic microwave background (CMB) radiation

When it was first proposed seriously in the late 1940s, many astronomers were not convinced by the Big Bang model. (They mostly preferred what was then known as the Steady State Theory of an unchanging universe.) However, the discovery in 1964 by Penzias and Wilson of **cosmic microwave background (CMB) radiation** provided the evidence that confirmed the Big Bang model for most astronomers. Penzias and Wilson discovered that low-level microwave radiation can be detected coming (almost) equally from all directions (it is **isotropic**), rather than from a specific source. (Later, important tiny variations were discovered in the CMB, a discovery that has vital implications for understanding the non-uniform structure of the universe and the formation of galaxies (Figure 16.43.)

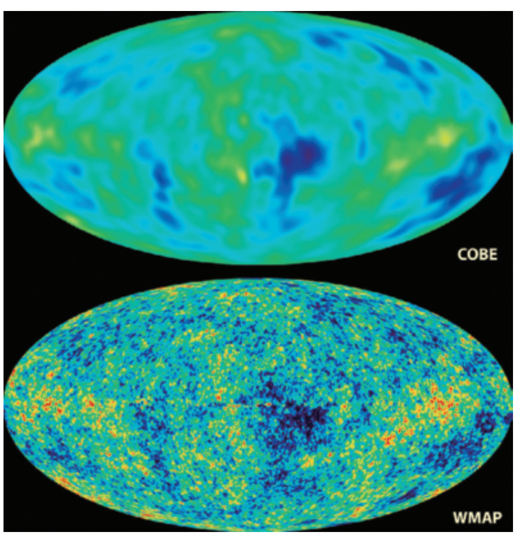


Figure 16.43 'Ripples in Space'; a map of the whole sky at microwave wavelengths showing the *very small* variations (1 part in 100 000) in the CMB radiation – firstly from COBE satellite and, later, in more detail from the WMAP satellite

Cosmic background radiation has been a major area of astronomical research for many years, including by the Cosmic Background Explorer (COBE) satellite. A very large amount of data has been collected and analysed by astronomers from many different countries.

We have seen before that everything emits electromagnetic radiation and that the range of wavelengths emitted depends on temperature. The Big Bang model predicts that the universe was incredibly hot at the beginning and has since been cooling down as it expands, so that the average temperature of the universe should now be about 2.76 K.

Figure 16.44 shows the black-body radiation spectrum emitted from matter at 2.76 K. When this isotropic radiation was discovered by Penzias and Wilson coming (almost) equally from all directions, the *Hot* Big Bang model was confirmed.

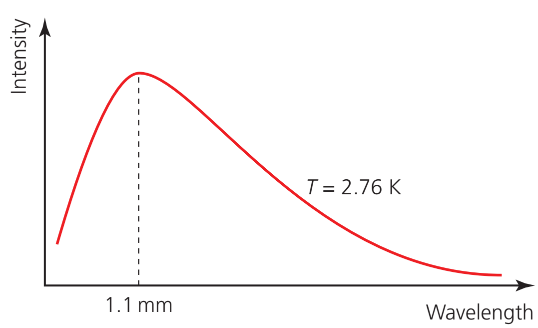


Figure 16.44 Spectral distribution for a temperature of 2.76 K

Wien's law (for black bodies) can be used to confirm the peak wavelength associated with this temperature:

$$\lambda_{max}T = 2.9 \times 10^{-3}\,\text{m K}$$

$$\lambda_{max} = \frac{2.9 \times 10^{-3}}{2.76} = 1.1 \times 10^{-3}\,\text{m}$$

An alternative (and equivalent) interpretation of CMB radiation is that the shorter wavelengths emitted when the universe was hotter have stretched out because of the expansion of space.

The observable universe

After the development of the Big Bang model it seemed that the universe could be finite, with a finite number of stars, each having a finite lifetime, thus limiting the amount of radiation that could reach Earth. More importantly, even if the universe is infinite, it was now known to have a definite age, which means that the universe that is observable to us is limited by the distance that light can travel in the time since the Big Bang.

The universe that we can (in theory) observe from Earth is a sphere around us of radius 4.6×10^{10} ly. This is known as the **observable universe** or the visible universe. (This distance is longer than 1.4×10^{10} ly because space has expanded since the Big Bang.) If there is anything further away, we cannot detect it because the radiation has not had enough time to reach us.

56 Summarize the two major discoveries that support the Big Bang model of the universe.

57 Astronomers look for 'shifts' in spectra as evidence for an expanding universe. The spectrum of which element is most commonly used, and why?

58 Draw a diagram to help explain why the light from some galaxies may be blue-shifted.

59 How will the average temperature of the universe change in the future if:
a the universe continues to expand
b the universe begins to contract?

The accelerating universe and red-shift (z)

What happens to the universe in the future is obviously dependent on the rate at which it is expanding and whether or not the expansion will continue indefinitely. Previously it was believed that the receding galaxies were simply losing kinetic energy and gaining gravitational potential energy, like objects projected away from the Earth, and that the fate of the universe depended on their initial speeds and the mass in the universe. But in recent years it has been discovered that the rate of expansion of the universe is *not* decreasing, but *increasing*. This is discussed in more detail in Section 16.5 (Additional Higher).

The evidence for the accelerating expansion of the universe comes from the observation of supernovae. When a certain kind (Type 1a) of supernova occurs, the energy released is always about the same and it is well-known to astronomers. This information can be used to determine the distance to such events using $b = L/4\pi d^2$. This means that such supernovae can be used as 'standard candles' for determining the distances to distant galaxies. Work on this topic by three physicists, Perlmutter, Riess and Schmidt (Figure 16.45), was jointly awarded the Nobel prize for physics in 2011.

Figure 16.45 Adam Riess, Saul Perlmutter and Brian Schmidt

The red-shifts from Type 1a supernovae have been found to be bigger than previously expected for stars at that distance away, strongly suggesting an 'accelerating universe'. This, of course, requires a new explanation and astronomers have proposed the existence of **dark energy**, a form of energy of low density, but present throughout the universe. Again, this will be discussed in more detail in Section 16.5.

Cosmic scale factor, *R*

Astronomers use the **cosmic scale factor** to represent the size of the universe by comparing the distance between any two specified places (two galaxies, for example) at different times. These distances, and the cosmic scale factor, increase with time because the universe is expanding.

$$\text{cosmic scale factor (at a time } t\text{)}, R = \frac{\text{separation of two galaxies at time } t}{\text{separation of the same two galaxies now}}$$

Because it is a ratio, the cosmic scale factor does not have a unit. It varies with time.

From the definition, it should be clear that at this time $R = 1$, in the past $R < 1$ and (in an expanding universe) in the future $R > 1$. If at some time in the future the universe has doubled in size, R will equal 2 at that time.

More generally, we can define the cosmic scale factor as follows:

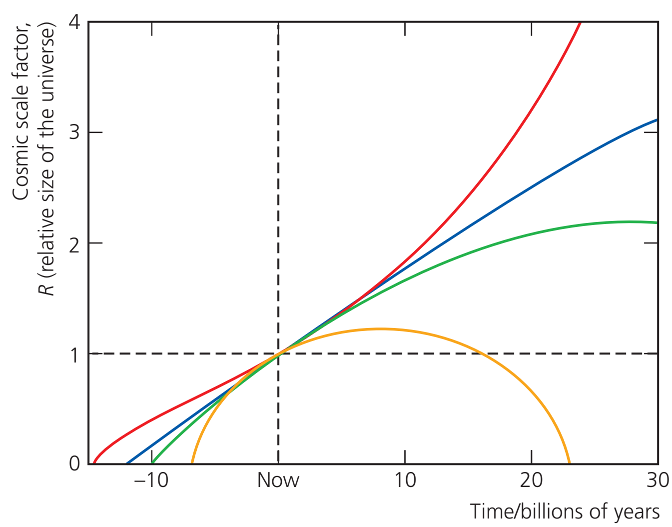
$$\text{cosmic scale factor (at a time, } t\text{)}, R = \frac{\text{separation of two galaxies at time } t}{\text{separation, } d_0\text{, of the same two galaxies at a specified time, } t_0}$$

$$R(t) = \frac{d(t)}{d_0}$$

Figure 16.46 shows some predictions for the possible size of the universe in the future (and how it might have been in the past).

- The red line represents an accelerating universe. This will be discussed in more detail in Section 16.5.
- The blue line represents a universe that will continue to expand for ever (but at a decreasing rate).
- The green line represents a universe that will expand for ever but at a rate that reduces to zero after infinite time.
- The orange line represents a universe that will reach a maximum size and then contract.

Figure 16.46 Possible futures for the universe



Relationship between red-shift and cosmic scale factor

We know that:

$$\text{red-shift, } z = \frac{\Delta\lambda}{\lambda_0} = \frac{\lambda - \lambda_0}{\lambda_0}$$

where λ is the wavelength received from a distant galaxy because of the expansion of space, and λ_0 is the wavelength that was emitted.

Because the expansion of the wavelength can be represented by an increase in the cosmic scale factor between the time the light was emitted, R_0, and the time it was received, R, we can write:

$$z = \frac{\lambda - \lambda_0}{\lambda_0} = \frac{R - R_0}{R_0}$$

or:

$$z = \frac{R}{R_0} - 1$$

This equation is given in the *Physics data booklet*.

Worked example

9 The light from a distant galaxy was found to have a red-shift of 0.16.
a What was the recession speed of the galaxy?
b Determine the cosmic scale factor when the light was emitted.
c Estimate the size of the observable universe at that time (size now = 4.6×10^{10} ly).

a $z \approx \frac{v}{c}$

$0.16 = \frac{v}{3.0 \times 10^8}$

$v = 4.8 \times 10^7\,\text{m s}^{-1}$

b $z = \left(\frac{R}{R_0}\right) - 1$

$0.16 = \left(\frac{1}{R_0}\right) - 1$

$R_0 = 0.86$

c $0.86 \times 4.6 \times 10^{10} = 4 \times 10^{10}$ ly

60 **a** Explain what is meant by the term 'standard candle'.
b Why are observations of supernovae considered to be the best way of determining the distances to remote galaxies?

61 Suggest what future for the universe is represented by the orange line in Figure 16.46.

62 Measurements of the light from a distant galaxy show that a line on its spectrum is 4.8×10^{-8} m longer than when measured on Earth. If the light was emitted with a wavelength of 6.6×10^{-7} m,
a what is the value of the red-shift?
b Calculate the cosmic scale factor at the time the light was emitted.

ToK Link

The history of astronomy has many paradigm shifts

A **paradigm** is a set of beliefs, or patterns of thought, with which individuals or societies organize their thinking about a particular issue, whether it is big or small. It is like a framework for all our thoughts and actions when, for example, we try to understand how electricity flows down a wire, or decide which foods are healthy to eat. In scientific terms, a paradigm could be said to be a pattern of beliefs and practices that effectively define a particular branch of science at any period of time. An obvious example from this chapter would be the set of ideas associated with the, now discredited, belief that the Earth is at the centre of the universe and the various consequences of that fundamental idea.

The phrase **paradigm shift** has been used increasingly during the last 50 years since it was first popularized by Thomas Kuhn and others in the early 1960s. It is used especially with respect to developments in science. There are plenty of examples which suggest that, while scientific understanding, knowledge and practices obviously evolve and, hopefully, improve over time, many of science and technology's greatest achievements have occurred following a relatively sudden (and perhaps unexpected or even seemingly unimportant) discovery or invention, or following the genius of an individual who has the insight to look at something in a completely new way. The phrase 'to think outside the box' has become very popular in recent years and it neatly summarizes an encouragement to look at a problem differently from the way others think about it (the 'box' being the paradigm).

A paradigm shift occurs when new insights, technology and discoveries have such a fundamental effect that current ideas or beliefs have to be rejected. Most individuals, organizations and societies find that a very difficult thing to do, even to the point where they strongly reject overwhelming evidence that their prevailing beliefs or actions are no longer reasonable. The response of the Roman Catholic Church to scientific evidence that the Earth revolved around the Sun was to simply ignore it and persecute those who held those beliefs.

The Big Bang model is another example of a paradigm shift in astronomical thinking and, if extra-terrestrial life were ever discovered, then most of us would look at ourselves in a completely new way – a tremendous paradigm shift! On a less profound scientific level, the technology of the internet and the introduction of social networking sites are having such a dramatic effect on the way that many people interact, that they can be described as producing a paradigm shift in communications. Characteristically, there are many people who are unwilling to accept such changes in their lives and who believe that they are unnecessary, or even harmful.

16.4 (D4: Additional Higher) Stellar processes –

the laws of nuclear physics applied to nuclear fusion processes inside stars determine the production of all the elements up to iron

We will begin by looking in more detail at the processes that lead to the formation of stars from **interstellar medium** (ISM), which is about 99% gas (mostly hydrogen and helium) and 1% dust. This medium has both very low temperature and low density. The medium will *not* be totally uniform (homogeneous) and it can also be disturbed by neighbouring stars or even the shock waves from a supernova.

The forces of gravity may be very small, but given enough time the gas and dust can gather together more where the density is slightly higher. The main thing stopping the eventual collapse of a nebula (or part of a nebula) under gravitational forces is the opposing pressure provided by the movement of the gas molecules.

Jeans criterion

The English astronomer Sir James Jeans did a lot of the early work on the conditions necessary for star formation. In simple terms, if the gravitational potential energy of a mass of gas is higher than the kinetic energy of its molecules, it will tend to collapse. A star cannot be formed unless the mass of the gas is greater than a certain critical value, called the **Jeans mass**, $\mathbf{M_J}$. The value of the Jeans mass is temperature dependent. If the interstellar medium is warmer, the mass necessary for the formation of a star will be higher.

The collapse of an interstellar cloud to form a star can only begin if its mass $M > M_J$.

Calculations using the Jeans criterion are *not* included in this course, but it is instructive to consider an example. For hydrogen gas with a tiny density of 10^8 atoms m^{-3} at a temperature of 100 K, the Jeans mass, M_J, is approximately 10^{33} kg. This mass is equivalent to at about 1000 × the mass of our Sun. These figures demonstrate that a very large mass is needed before interstellar medium can begin to collapse. Although when collapse does occur, the mass involved is large enough for the formation of more than one star.

Nuclear fusion

Nuclear fusion in the cores of stars is the dominant energy transfer that provides the power for the radiation they emit. Fusion can occur because of the very high temperatures created when gravitational potential energy is transferred to the kinetic energy of particles as the interstellar mass has been pulled together by gravitational forces.

It can take a long time for stars to fuse most of their hydrogen into helium and during this time they are known as main sequence stars. When the supply of hydrogen in a star begins to be depleted ('run out'), different fusion processes can occur (at higher temperatures) as the star enters the later stages of its lifetime and moves off the main sequence on the HR diagram. We will now examine these fusion processes in greater detail.

Nuclear fusion in the main sequence

Earlier in this chapter we summarized the fusion of hydrogen into helium:

$$4{}^{1}_{1}\text{H} \rightarrow {}^{4}_{2}\text{He} + 2{}^{0}_{1}\text{e} + \text{neutrinos and photons}$$

But the process is a little more complicated. It is known as the **proton–proton cycle** and has three separate stages:

- Two protons fuse to make a ${}^{2}_{1}$H (deuterium) nucleus. In this process a positron and an (electron) neutrino are emitted.

$${}^{1}_{1}\text{H} + {}^{1}_{1}\text{H} \rightarrow {}^{2}_{1}\text{H} + {}^{0}_{1}\text{e}^{+} + {}^{0}_{0}\nu_{e}$$

- The deuterium nucleus fuses with another proton to make He-3. In this process a gamma ray photon is emitted.

$${}^{2}_{1}\text{H} + {}^{1}_{1}\text{H} \rightarrow {}^{3}_{2}\text{He} + {}^{0}_{0}\gamma$$

- Two He-3 nuclei combine to make He-4. Two protons are released in this reaction.

$${}^{3}_{2}\text{He} + {}^{3}_{2}\text{He} \rightarrow {}^{4}_{2}\text{He} + 2{}^{1}_{1}\text{H}$$

These three stages are illustrated in Figure 16.47.

Figure 16.47 The proton–proton cycle

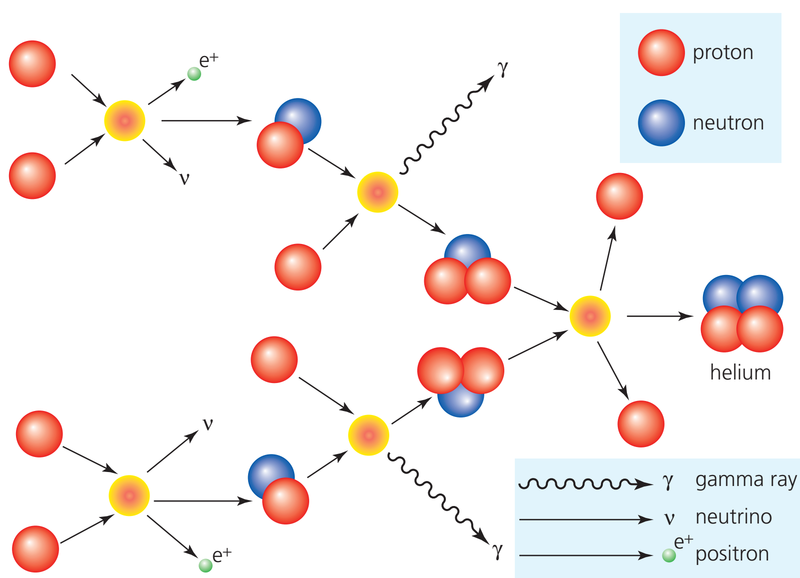
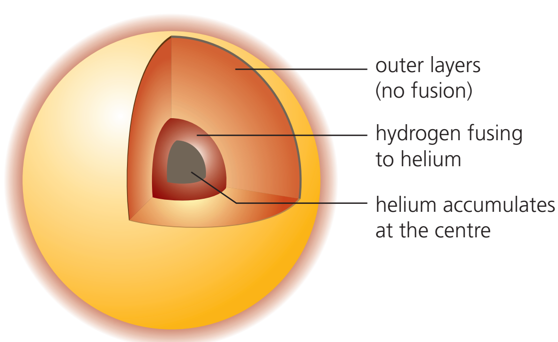


Figure 16.48 Internal structure of a main sequence star



The nuclear potential energy released in each cycle is 26.7 MeV and this is transferred to the kinetic energy and electromagnetic energy of the products. Energy is continually transferred to the surface of the star, from where it is radiated away at the same rate as it is produced by nuclear fusion, so the star remains in equilibrium. Because helium atoms are more massive than hydrogen atoms, they will remain near the centre of the star where they were formed (see Figure 16.48).

Main sequence stars remain stable for a long time because thermal gas pressure and radiation pressure outwards oppose the gravitational pressure inwards.

Time for which stars stay on the main sequence

When the supply of hydrogen becomes sufficiently depleted (after about 12 % of the total mass of hydrogen in the star has been fused) the star will no longer be in equilibrium. The inert helium core will begin to collapse inwards under the effect of gravitational forces and this marks the beginning of its end as a main sequence star. The 'lifetime' of the star as a main sequence star depends on the original mass of hydrogen and the rate of nuclear fusion. But more massive stars have more concentrated cores at higher temperatures and this means that they deplete their hydrogen *much* quicker.

More massive stars have shorter main sequence lifetimes.

Earlier in this chapter we introduced the following equation linking the luminosity, L, of a main sequence star to its original mass, M:

$$L \propto M^{3.5}$$

This equation is given in the *Physics data booklet*.
The luminosity of a stable main sequence star can be assumed to be constant, so:

$$L = \frac{\text{total energy released by nuclear fusion}}{\text{time spent as a main sequence star, } T}$$

It is reasonable to assume that the energy released is approximately proportional to the mass of the star, so:

$$L \propto \frac{M}{T}$$

Combining these last two equations, we get:

$$M^{3.5} \propto \frac{M}{T}$$

or:

$$T \propto \frac{1}{M^{2.5}}$$

This equation is *not* given in the *Physics data booklet*.

The lifetime of our Sun

We know the following facts about the Sun:

- mass, $M_{\odot} = 1.99 \times 10^{30}\,\text{kg}$
- luminosity = $3.85 \times 10^{26}\,\text{W}$
- Each proton–proton cycle releases 26.7 MeV (= 4.27×10^{-12} J) of energy.
- When they are first formed, main sequence stars consist of approximately 75% hydrogen.
- It will end its main sequence lifetime when about 12% of its hydrogen has been fused into helium.

We can calculate a value for its main sequence lifetime as follows:

amount of hydrogen that will be fused ('burned') during the main sequence lifetime = 12% of 75% of $1.99 \times 10^{30}\,\text{kg} = 1.79 \times 10^{29}\,\text{kg}$

mass involved with each proton–proton cycle = $4 \times 1.67 \times 10^{-27} = 6.68 \times 10^{-27}\,\text{kg}$

number of proton–proton cycles during the main sequence lifetime

$$= \frac{1.79 \times 10^{29}}{6.68 \times 10^{-27}} = 2.68 \times 10^{55}$$

current rate of proton–proton cycles $= \dfrac{3.85 \times 10^{26}}{4.27 \times 10^{-12}} = 9.02 \times 10^{37}\,\text{s}^{-1}$ (assumed constant for the main sequence lifetime)

main sequence lifetime of our Sun $= \dfrac{2.68 \times 10^{55}}{9.02 \times 10^{37}} = 2.97 \times 10^{17}\,\text{s}$ (or about 9.4×10^9 years)

Our Sun (mass $M_\odot$) is expected to stay on the main sequence a total of about 10^{10} years.

We can also estimate the decrease in the mass of the Sun due to nuclear fusion reactions from $\Delta E = \Delta mc^2$ (Chapter 7):

energy transferred during main sequence lifetime, ΔE = power × time

$= 3.85 \times 10^{26} \times 2.97 \times 10^{17} = 1.14 \times 10^{44}\,\text{J}$

$\Delta E = \Delta mc^2$

$1.14 \times 10^{44} = \Delta m \times (3.0 \times 10^8)^2$

$\Delta m = 1.3 \times 10^{27}\,\text{kg}$

This is equivalent to $4.3 \times 10^9\,\text{kg}\,\text{s}^{-1}$!

Worked example

10 Estimate the main sequence lifetime of a star that is twice the mass of our Sun.

$$\frac{T_{\text{star}}}{T_{\text{Sun}}} = \left(\frac{M_\odot}{M_{star}}\right)^{2.5}$$

$$= \left(\frac{1}{2}\right)^{2.5}$$

$$\log \frac{T_{\text{star}}}{T_{\text{Sun}}} = 2.5 \log 0.5 = -0.753$$

$$\frac{T_{\text{star}}}{T_{\text{Sun}}} = 0.17$$

$$T_{\text{star}} = 0.17 \times T_{\text{Sun}}(10^{10}) = 1.7 \times 10^9 \text{ years}$$

The star is twice as massive, but its lifetime is shorter than $\frac{1}{5}$ that of the Sun.

63 An approximate value for the Jeans mass can be calculated from the following equation (which does not need to be remembered):

$$M_J^2 = \left(\frac{5kT}{Gm}\right)^3 \times \frac{3}{4\pi\rho}$$

where M is the mass of the individual atoms particles and ρ is the mean density of the medium.

a Calculate a value for M_J for hydrogen at a temperature of 40 K and density of 200 molecules per cm^3.

b i Explain why a greater Jeans mass would be needed for the same gas with the same density, but at a higher temperature.

ii If the temperature was 50 K instead of 40 K (at the same density), by what factor would the Jeans mass increase?

64 a Estimate the luminosity of a star (in terms of $L_\odot$) which has a mass ten times heavier than the mass of the Sun.

b What is the approximate mass of a star (in terms of $M_\odot$) which is half as luminous as the Sun?

65 a What mass of star (in terms of $M_\odot$) will have a main sequence lifetime twice as long as the Sun?

b Estimate the lifetime of a star that has a mass 20 times greater than our Sun.

66 a A main sequence star has a luminosity of 4.9×10^{28} W. What is the annual decrease in mass due to nuclear fusion?

b What is this change of mass expressed as a percentage of the original mass of the star?

Nucleosynthesis off the main sequence

After the hydrogen becomes depleted, the core of the star begins to contract because, once the rate of fusion is reduced, the gravitational forces are greater than the outward forces. Gravitational potential energy is then transferred to kinetic energy of the nuclei in the core, and to increased thermal gas pressure, which results in the significant expansion of the outer layers of the star. This results in the envelope cooling, creating a red giant or red supergiant (as described earlier in this chapter). These changes cause the star to leave the main sequence.

The temperatures in the cores of red giants are sufficient to cause the fusion of helium nuclei (or heavier elements). The creation of the nuclei of heavier elements by fusion is called **nucleosynthesis**.

In general, the contraction of the cores of main sequence stars of greater mass will result in higher temperatures, which means that the nuclei then have higher kinetic energies, so that they can overcome the larger electric repulsive forces involved in the fusion of heavier elements.

The following is a simplified outline of the processes involved:

- For stellar masses lower than $4M_{\odot}$ (red giants), the core temperature can reach 10^8 K and this is large enough for the nucleosynthesis of carbon and oxygen. (Helium is still produced in an outer layer.) For example:

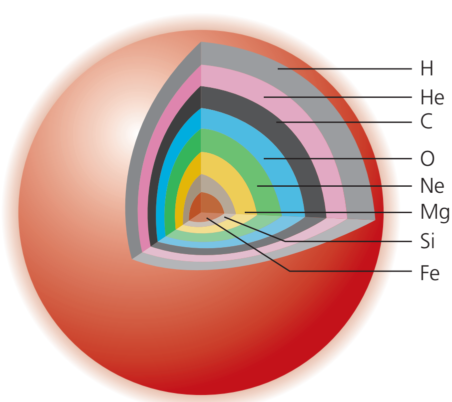
$$^4_2\text{He} + ^4_2\text{He} \rightarrow ^8_4\text{Be} + ^0_0\gamma$$

$$^8_4\text{Be} + ^4_2\text{He} \rightarrow ^{12}_6\text{C} + ^0_0\gamma$$

$$^{12}_6\text{C} + ^4_2\text{He} \rightarrow ^{16}_8\text{O} + ^0_0\gamma$$

- For stellar masses between $4M_{\odot}$ and $8M_{\odot}$ (larger red giants), the core temperature exceeds 10^9 K and this is large enough for the nucleosynthesis of neon and magnesium. (Helium, carbon and oxygen are still produced in the outer layers.)
- For stellar masses over $8M_{\odot}$ (red supergiants), the core temperature is large enough for the nucleosynthesis of elements as heavy as silicon and iron. (The lighter elements are still produced in the outer layers.) From Chapter 7 we know that the nucleus of iron is one of the most stable (it has one of the highest average binding energies per nucleon). This means that there would have to be an energy input to create heavier nuclei. (See later in this chapter.)

The structure of stars off the main sequence is layered ('like the skins of an onion') as the heavier elements are found closer to the centre. A red supergiant will have the most layers (see Figure 16.49).



■ **Figure 16.49** The layers of a red supergiant

67 Write a possible equation for a nuclear reaction which produces silicon-14.

68 What elements will be in the Sun when it finishes its main sequence lifetime?

69 Explain why nuclear fusion within main-sequence stars cannot produce nuclides with nucleon numbers greater than 62.

The formation of elements heavier than iron by neutron capture

Neutrons are produced in some nuclear fusion reactions within a star. For example:

$$^{13}_6\text{C} + ^4_2\text{He} \rightarrow ^{16}_8\text{O} + ^1_0\text{n}$$

Because neutrons are uncharged, they do not experience electrostatic repulsion and they can get close enough to nuclei to come within range of the attractive nuclear strong forces and be 'captured'.

$$^1_0\text{n} + ^A_Z\text{X} \rightarrow ^{A+1}_{Z}\text{X}$$

If this happens, it will increase the mass number of the nucleus and affect its stability. The nucleus may then decay by the emission of a beta negative particle, which will result in the formation of a new element with a higher proton number (Chapter 7). But various other outcomes are also possible, depending on factors such as the neutron density, the temperature and the half-life of the beta decay. We can identify two principal processes: the s-process and the r-process.

The s-process: slow neutron capture

The **s-process** occurs in certain kinds of red giants over a long time at relatively low neutron density and intermediate stellar temperatures. Under these conditions neutron capture is much less probable than beta decay, which means that after neutron capture, beta decay will nearly always occur before the capture of another neutron. In general:

$$ {}^{1}_{0}n + {}^{A}_{Z}X \rightarrow {}^{A+1}_{Z}X \rightarrow {}^{A+1}_{Z+1}Y + {}^{0}_{-1}e + \overline{\nu} $$

In this way, starting from the heavier nuclides found in a red giant star, heavier and heavier elements can be created over a very long time, but none heavier than bismuth-209.

The r-process: rapid neutron capture

This process can form nuclides of a wide range of elements, including the most massive. The **r-process** occurs very quickly in supernovae, which have very high neutron densities and temperatures. Under these conditions, neutron capture is much more likely than beta decay, which means that repeated neutron captures can rapidly result in nuclides with large mass numbers. Beta decay causes transmutation to different elements and this process is assisted by the presence of large numbers of neutrinos in a supernova.

Supernovae

Supernovae are sudden, unpredictable and very luminous stellar explosions. (The name 'supernova' expresses these facts: super-bright and new.) For a few weeks their luminosities can be higher than an entire galaxy. These unusual events have time scales that are very different from most other events in the universe and this makes them both fascinating and useful to astronomers.

It has been estimated that a supernova will occur about once in every 50 years in a typical galaxy. The last known supernova in our own galaxy was visible without a telescope and occurred in 1604 (Figure 16.50). Its appearance was used by astronomers of the time as evidence that the universe was *not* fixed and unchanging. It occurred about 20 000 ly from Earth and was visible in the daytime for a few weeks. Its remnant can still be seen with a telescope. Until recently the observation of supernovae was fairly random and many were detected by amateur astronomers, but now the search has become much more automated and computer controlled.

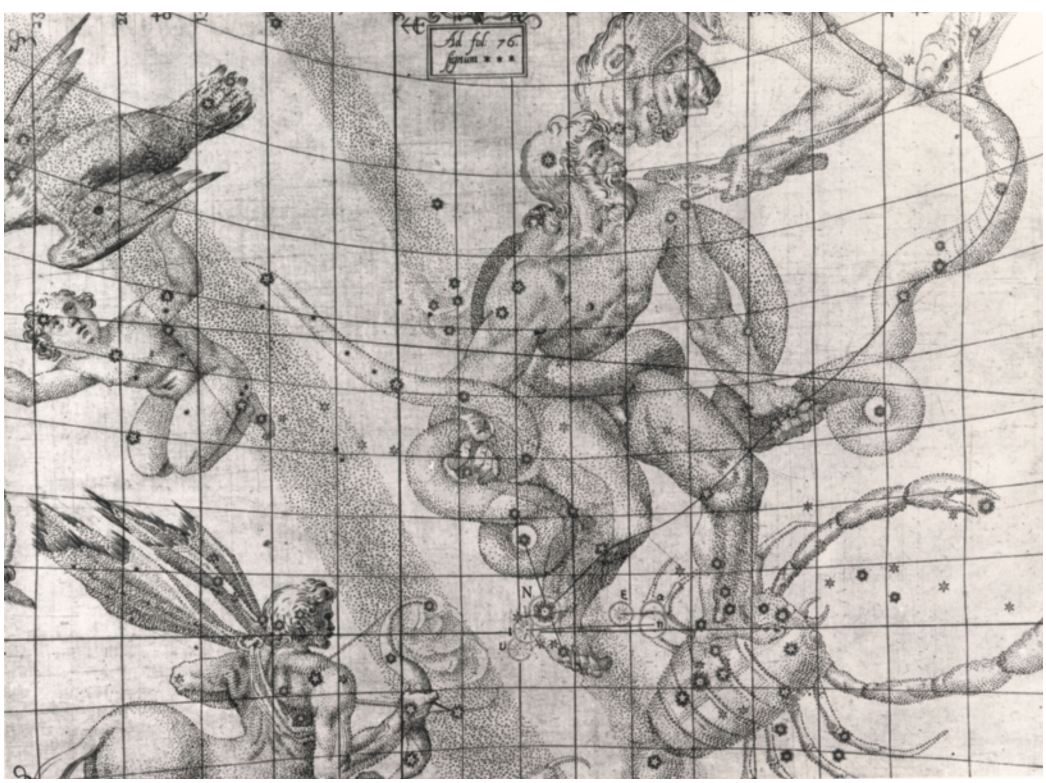


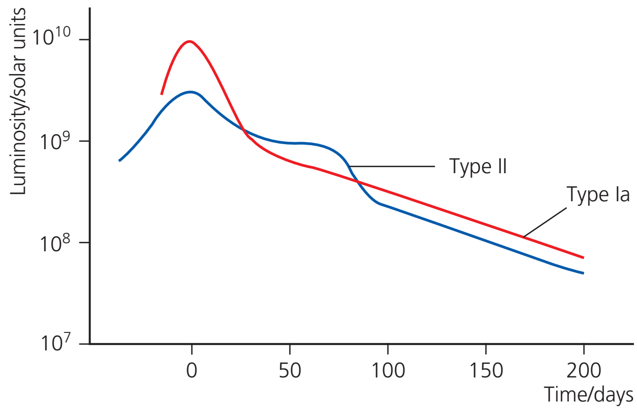
Figure 16.50 Johannes Kepler's original drawing showing the 1604 supernova in the constellation of Ophiuchus (the 'serpent bearer'). It is shown with the letter *N* which is nine grid squares down from the top and eight from the left

Supernovae are important in the creation of heavy elements and they can also be responsible for causing disturbances in the interstellar medium that can instigate the birth of stars.

Type Ia and Type II supernovae

Supernovae are classified initially by reference to the hydrogen lines in their spectra but also from their 'light curves' (see Figure 16.51).

Figure 16.51 Light curves for two types of supernovae



Type Ia supernovae

If one of the stars in a binary system has become a white dwarf (with a carbon–oxygen core), its gravitational field may be strong enough to attract matter from its neighbouring star. This increase in mass means that electron degeneracy pressure is no longer high enough to resist the collapse of the star. (The mass of the white dwarf has risen to the Chandrasekhar limit for the maximum mass of a white dwarf (= $1.4M_\odot$).) The rapid increase in temperature resulting from the transfer of gravitational potential energy to the kinetic energy of particles causes rapid and widespread carbon fusion, producing a supernova which, in its early stages, could be 10^{10} more luminous than the Sun (see Figure 16.52).

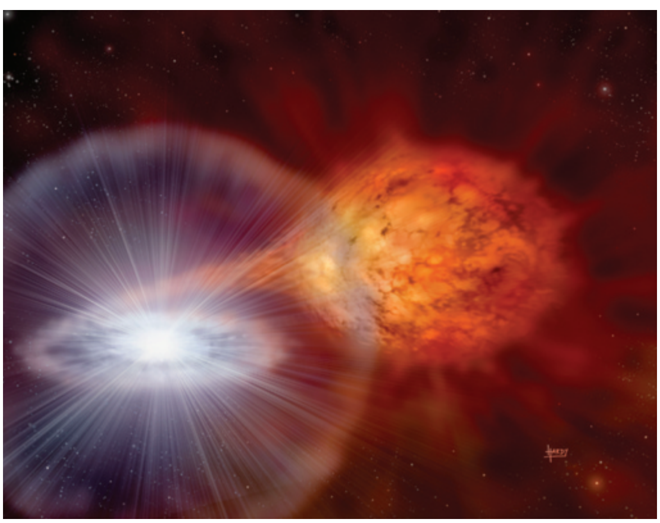


Figure 16.52 An artist's impression of a Type Ia supernova

Because these reactions only occur when the star has acquired a certain (well-known) mass, the luminosities, L, of Type Ia supernovae are always about the same. This means that the distance, d, from a Type Ia supernova to Earth can be determined using $b = L/4\pi d^2$, once the initial apparent brightness, b, has been measured. As mentioned earlier in the chapter, this has made this type of supernova very useful in determining the distance to galaxies in which supernovae occur. This is why they are known as 'standard candles' for galaxies up to 1000 Mpc away from Earth.

Type II supernovae

When the nuclear reactions in a red supergiant finish, the star collapses but the mass and energy involved are so huge that the nuclei in the core get deconstructed back to protons, neutrons, electrons, photons and a large number of neutrinos. The process of these particles interacting is complicated, but the consequence is an enormous shock wave travelling outwards, tearing apart the outer layers of the star and spreading enormous distances into the surrounding space. Rapid neutron capture occurs, resulting in the creation of the heavier elements. The remaining core will become a neutron star or a black hole (as discussed earlier in the chapter).

Additional Perspectives

'Where did the atoms in our bodies come from?

This question can have different answers depending on the timescale being considered. Of course, chemical and biological processes are responsible for redistributing atoms and it is not difficult to believe that a carbon atom in a person's body was part of a plant growing on the other side of the Earth or a fish swimming in a river six months earlier. Looking further back, the same atom might have been part of a dinosaur millions of years ago.

It is unlikely that any of the atoms that were in your body when you were born are still in your body. Over a long period of time it may be reasonable to assume that atoms are redistributed randomly, and a famous kinetic theory question asks students to estimate, for example, how many of the carbon atoms that were in the body of Isaac Newton are now part of their body.

On the cosmological scale, which covers billions of years, all the atoms in your body were originally in interstellar space, nebulae, stars and supernovae, and their origins go further back to the earliest stages of the universe and the Big Bang. We are truly made of 'stardust'. And looking forward billions of years, that is the state to which our atoms will return.

1 Find out how long a typical atom remains in a human body.

ToK Link

A philosophical question

If our bodies and brains are no longer made of the same atoms that they were before, to what extent are we the same person?

70 Write down another possible nucleosynthesis reaction that results in the emission of a neutron.

71 Write down equations to represent the following s-process: iron-56 captures three neutrons and then emits a beta-negative particle and transmutes into cobalt.

72 Traces of uranium are found on Earth in many places. Explain where they came from.

73 Summarize the differences between Type Ia and Type II supernovae.

16.5 (D5: Additional Higher) Further cosmology – *the modern field of cosmology uses advanced experimental and observational techniques to collect data with an unprecedented degree of precision, and as a result very surprising and detailed conclusions about the structure of the universe have been reached*

Astronomical and cosmological research has been a substantial growth area in science in recent years. Advances in satellite engineering, imaging techniques and the computerized collection and analysis of data in many countries have all contributed to a rapid growth in our knowledge of the universe which, in turn, has increased the relative importance of astronomy in the spectrum of scientific subjects and also raised the general public's awareness and interest.

The knowledge base of an astronomer must now extend from the very large (obviously) to the very small, because the properties of sub-atomic particles and radiation are fundamental to the behaviour of stars and understanding the beginnings of the universe.

In this section we will discuss the central issues of the latest cosmological research – dark matter, dark energy and the possible fate of the universe. But we will begin by explaining the *cosmological principle* and clarifying our understanding of red-shifts.

The cosmological principle

This is a starting point in developing an understanding of the universe. It has already been implied throughout this chapter, but it should be stated formally. Astronomers believe that the *large-scale* structure of the universe is the same everywhere and that when we look in different directions, we see essentially the same thing.

It may seem that this chapter has involved the discussion of astronomical differences, for example, between different types of stars. In addition, when we look at the night sky, the views in different directions may be similar, but they are obviously not exactly the same. However, the universe is unimaginably enormous and on that scale these differences are usually insignificant.

The **cosmological principle** can be summarized as:

The universe is homogeneous and isotropic.

- The universe is **homogeneous** – any large section of the universe will be similar to any other large section.
- The universe is **isotropic** – what can be observed by looking in any one direction from anywhere in the universe is similar to what can be seen by looking in any other direction from the same place, or in any direction from any other place.

It may seem that if the first point is accepted, then the second must be true, but if the universe had an 'edge', then for some observers the view could be different in different directions, even if the universe was homogeneous.

Isotropy implies that the universe has no edge and no centre.

The cosmological origin of red-shift

When electromagnetic radiation is received that has a longer wavelength than when it was emitted, it is described as having a red-shift. There is more than one possible reason why radiation may be red-shifted.

1. The source of radiation and the observer could be moving apart from each other in unchanging space. This is known as the **Doppler effect**. The radiation will be shifted towards shorter wavelengths if the separation is decreasing (blue-shift). The change in wavelength can be determined from the equation $\frac{\Delta\lambda}{\lambda_0} \approx v/c$ but only if $v \ll c$ (maybe for $v = 0.25c$ and below).
2. The space between the source and the observer has expanded between the time when the radiation was emitted and the time when it was received. The wavelengths are increased by the same factor as the space. This is known as the **cosmological red-shift**. Cosmological blue-shifts are not possible.

The cosmological red-shift of radiation from distant galaxies provides evidence for the Big Bang model of the universe.

The equation $\Delta\lambda/\lambda_0 \approx \frac{v}{c}$ (if $v \ll c$) can also be applied to cosmological red-shifts (but for the most distant galaxies, this simplified equation cannot be used).

3. (**Gravitational red-shifts** (discussed in Chapter 13) occur as a result of radiation moving out of gravitational fields.)

It is possible for radiation from a star or galaxy to have *both* a cosmological red-shift and a Doppler shift. For example, some binary star systems receding from Earth will have a cosmological red-shift, while the two stars have Doppler shifts (one blue and one red). All galaxies will tend to recede, but a 'nearby' galaxy in a cluster will have a relatively low recession speed and, therefore, may be moving towards Earth in its orbit within the cluster, such that the Doppler blue-shift exceeds the cosmological red-shift.

A detailed calculation of any astronomical red-shift would need to take all these factors into account. This is *not* required for this course.

Mass in the universe

We can consider that the galaxies are gaining gravitational potential energy and losing kinetic energy as space expands and their separations increase. If the galaxies have sufficient energy, this expansion can continue forever; if they do not have enough energy, they will eventually be pulled back together by gravitational forces and space will contract. We know the recession speeds of galaxies, so in principle it would be a straightforward calculation using classical gravitation theory to determine what will happen in the future, *if* we have accurate information about the mass in the universe.

Astronomers commonly refer to the *average density* (total mass/total volume) of the universe.

Critical density

The **critical density**, ρ_c, of the universe is the theoretical density that would *just* stop the expansion of the universe after an infinite time.

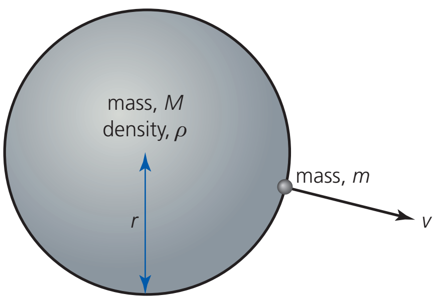


Figure 16.53 Estimating critical density

A theoretical value for the critical density can be obtained using classical Newtonian gravitation, as follows.

Consider a homogeneous spherical cloud of interstellar matter of mass M, radius r and density ρ (see Figure 16.53). A mass m at a distance r from the centre is moving away with a speed $v = Hr$, where H is the value for the Hubble 'constant' at that particular time.

The total energy, E_T, of mass m is equal to the sum of its kinetic energy and gravitational potential energy:

$$E_T = \frac{1}{2}mv^2 + (-GMm/r)$$

remembering that gravitational potential energy is negative.

But $M = \frac{4}{3}\pi r^3 \rho$ and $v = Hr$, so:

$$E_T = \frac{1}{2}m(Hr)^2 - G\frac{4}{3}\pi r^3 \rho m/r$$

If mass m moves outwards until it reaches infinity after an infinite time, its total energy will then be zero. (This is similar to the calculation of an escape velocity from a planet, covered in Chapter 10). If $E_T = 0$ and $\rho = \rho_c$:

$$\frac{1}{2}m(Hr)^2 - G\frac{4}{3}\pi r^3 \rho_c m/r = 0$$

Simplifying, we get:

$$\rho_c = \frac{3H^2}{8\pi G}$$

This equation is given in the *Physics data booklet*.

Worked example

11 Calculate a value for the current critical density of the universe using $H = H_0 = 70\,\text{km s}^{-1}\,\text{Mpc}^{-1}$.

In SI units:

$$H_0 = \frac{70 \times 10^3}{3.26 \times 10^6 \times 9.46 \times 10^{15}} = 2.27 \times 10^{-18}\,\text{s}$$

$$\rho_c = \frac{3H^2}{8\pi G} = \frac{3 \times (2.27 \times 10^{-18})^2}{8\pi \times 6.67 \times 10^{-11}} = 9.2 \times 10^{-27}\,\text{kg m}^{-3}$$

The value of the critical density is time dependent, decreasing as the universe expands (and H changes). The current *theoretical* value is believed to be the equivalent of about (only) six hydrogen atoms per cubic metre.

74 **a** Explain why looking at the night sky on a clear night you can see many more stars in some directions than in others.
b Explain why this fact does not contradict the cosmological principle.

75 Explain why cosmological blue-shifts are never detected.

76 Suggest why high-energy physics experiments at CERN may help astronomers understand the early universe.

77 Explain why the equation $\Delta\lambda/\lambda_0 \approx v/c$ cannot be used to calculate the recession speed of a very distant galaxy accurately.

78 Suggest why astronomers refer to density of the universe, rather than its mass.

79 **a** Show that a density of $9.2 \times 10^{-27}\,kg\,m^{-3}$ is approximately equivalent to six hydrogen atoms per cubic metre.
b Compare the critical density of the universe to a typical density for the air you are breathing.

Estimating the actual average density of the universe

What happens to the universe in the future depends on how the *actual* average density compares to the critical density. Because we have a good understanding of the masses of individual stars, galaxies and nebulae, as well as their separations and distribution, it would seem that straightforward calculations should lead to an estimate of the total mass in a typical large volume, from which an average density can be calculated. This is certainly possible, but the results of such calculations applied to galaxies are inaccurate and they are much less than the more reliable values of mass determined by applying the laws of physics to the rotation of galaxies.

Rotation curves and the mass of galaxies

The theoretical velocities of stars rotating with a galaxy around its centre of mass can be calculated from the laws of classical physics. It is convenient to consider the stars close to the centre separately from the more distant stars, as follows.

1 *Close to the centre of a galaxy* the circular orbital velocity of a star increases approximately proportionately to its distance from the centre. We can explain this using Newtonian gravitation: assume that a galaxy is homogeneous and spherical, with radius r and density ρ. It would act as if all its mass, M, was concentrated at its centre. A mass m on the circumference would experience a force GMm/r^2. Equating this to the expression for centripetal force we get:

$$\frac{mv^2}{r} = \frac{GMm}{r^2}$$

so that:

$$v^2 = \frac{GM}{r}$$

Because $M = \frac{4}{3}\pi r^3 \rho$:

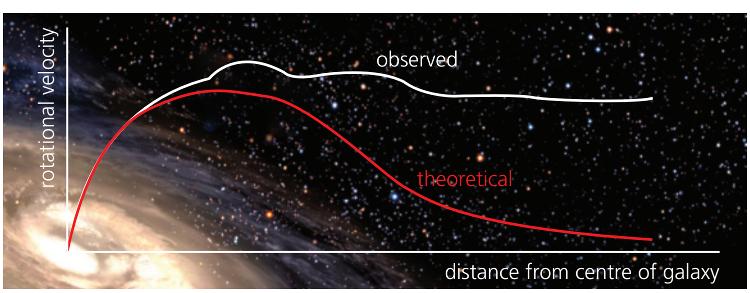
$$v^2 = \frac{4}{3}\pi G r^2 \rho$$

or:

$$v = \sqrt{\frac{4\pi G\rho}{3}}\, r$$

This equation, showing that the rotational speed of the galaxy is proportional to the distance from the centre, is given in the *Physics data booklet*. Because of the simplifying assumptions involved, it only gives an approximate value for the speeds closer to the centre of the galaxy.

2 *At longer distances* the situation becomes similar to, for example, planets freely orbiting a massive central star, so that the rotational velocity decreases with distance.
($v \propto \sqrt{\frac{1}{r}}$ – see Chapter 6)



■ **Figure 16.54** Comparison of theoretical and observed rotational curves

Detailed analysis will produce a combined theoretical **rotational curve** (graph of v–r), such as shown in red in Figure 16.54. However, actual measurements made from red-shift observations show a significantly different pattern (shown in white in Figure 16.54). (The red-shifts due to the recession of the galaxy as a whole and the orbital motions of the stars need to be distinguished.) The calculated orbital speeds for the outer stars of the galaxy are approximately constant.

The difference between theory and observation demands an explanation. It could be that there is something wrong with the basic physics used in the theory, but the preferred explanation is that there must be a lot more mass in galaxies than can be observed directly, and that this mass is concentrated in the outer reaches of the galaxy in a **dark matter halo**.

Dark matter

Dark matter is the name given to the proposed matter that must be present in the universe, but which has never been detected directly because it neither emits nor absorbs radiation.

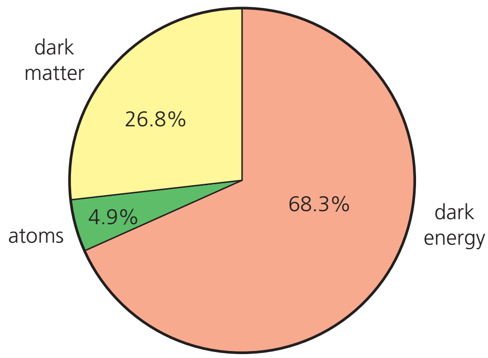


Figure 16.55 The constituents of the universe

Dark matter is believed to be about five times more plentiful in the universe than observable matter (atoms), but **dark energy** (see later in this section) is thought to be the dominant form of mass–energy in the universe (see Figure 16.55).

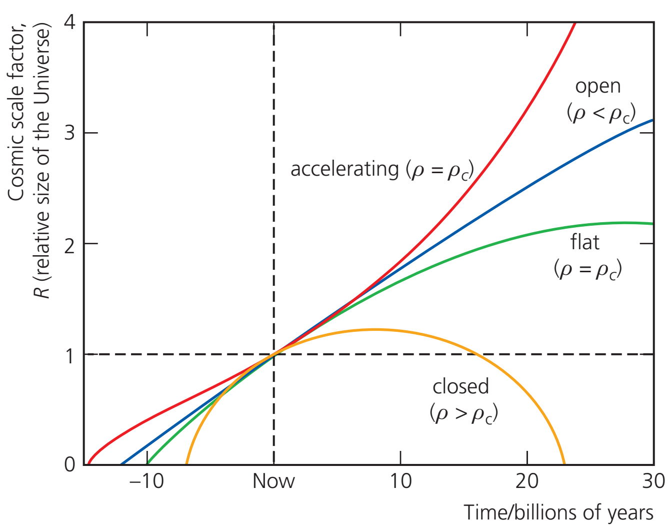
Many theories have been proposed about the nature of dark matter, including:

- **MACHOs** – **m**assive **a**stronomical **c**ompact **h**alo **o**bjects. These could be old stars or small 'failed' stars, very large planets or even black holes, any or all of which simply do not emit enough radiation for us to detect them. It is considered unlikely that there could be a sufficient number of such objects to explain dark matter fully.
- **WIMPs** – **w**eakly **i**nteracting **m**assive **p**articles. (Note that here 'massive' simply means 'with mass'; it does not imply large individual masses.) There could be very large numbers of particles that we do not know about yet simply because they are very difficult to detect. Including…
- **Neutrinos** – the masses of neutrinos are presumed to be very small, but they are still not known with any certainty. Neutrinos are present in the universe in vast numbers so they could contribute significant mass.

Possible futures for the universe

Figure 16.56 shows possible futures for the universe as it expands. This diagram is similar to Figure 16.46 but it is worth repeating here because it is so important. Note that the density affects not only the future, but also our estimate of the age of the universe.

Figure 16.56 How the average density of the universe affects its future



- **A flat universe** ($\rho = \rho_c$) – the rate of expansion will reduce to zero after an infinite time.
- **A closed universe** ($\rho > \rho_c$) – if the density of the universe is higher than the critical density, then at some time in the future the universe will stop expanding and then begin to contract and eventually end as a 'Big Crunch'.
- **An open universe** ($\rho < \rho_c$) – if the density of the universe is lower than the critical density, then the universe will continue to expand forever.

An accelerating universe

The latest calculations (involving dark matter) suggest that the actual density of the universe is (surprisingly?) very close to the critical density, which might suggest that any of the three possibilities listed above are feasible. However, as mentioned earlier in this chapter, recent measurements on the red-shifts of galaxies containing Type Ia supernovae indicate that the expansion of the universe has been *accelerating* (for about half its lifetime), as shown by the red line in Figure 16.56. This requires a different explanation: **dark energy**.

Nature of Science

Cognitive bias

In science, as in all areas of human experience, we usually believe what we want to believe. In other words, we are inclined to accept information that supports our existing experiences or beliefs, and inclined to reject information that would cause us to change our thinking. Of course, scientists strive to be objective and evaluate new data or theories without bias, but **confirmation bias** is present, often without people being aware of it, and this tends to make people reject the unexpected in favour of accepting the expected.

The concept that the expansion of the universe was slowing down because of the gravitational attraction between galaxies was widely accepted and in line with available evidence. The recent paradigm shift to an accelerating universe (especially without any substantial supporting explanation) has only been accepted because of the overwhelming evidence.

ToK Link

Limits of understanding

Experimental facts show that the expansion of the universe is accelerating yet no one understands why. Is this an example of something that we will never know?

British Astronomer Royal Martin Rees is quoted as saying:

'A chimpanzee can't understand quantum mechanics. It's not that the chimpanzee is struggling to understand quantum mechanics. It's not even aware of it. There is no reason to believe that our brains are matched to understanding every level of reality.'

While it may be reasonable not to worry about things that we do not know, and of which we are totally unaware, it is unlikely that most scientists would ever happily accept that something as important as the fate of the universe can never be known, especially when it is a central theme of modern astronomical research. Similar comments may apply to the reasons for the origin of the universe: we may never know, but astronomers will keep looking for answers.

Dark energy

Dark energy has been proposed in the last 20 years as an explanation for the accelerating expansion of the universe.

Dark energy has been described as providing a 'negative pressure' on the universe; a pressure opposing any possible contraction. It has not been detected directly and is still hypothetical, although it is believed that it may be homogeneous – throughout all space at a *very* low density. Even at such a low density, because it is everywhere, it may provide about 68% of the mass–energy content of the universe (see Figure 16.55). Many astronomers consider that it is an intrinsic property of space and that when space expands, so too will the amount of dark energy.

80 Use data available in this chapter (or elsewhere) to make an approximate order of magnitude estimate for the total mass of all the stars in the observable universe.

81 Use the internet to learn more about WIMPS and MACHOS.

82 The diameter of the observable Universe is 28×10^9 pc. Estimate its total mass.

83 If the Hubble constant was reassessed to be 10% higher, by what factor would the value of critical density of the universe change?

84 Use classical gravitational theory (as used in Chapter 10 for satellite orbits) to show that we might expect the orbital speeds of stars a long way from the galactic centre to be inversely proportional to the square root of their distance from the centre.

85 The maximum Doppler shift, z, due to a star orbiting within its galaxy at a distance of 9.5×10^{15} km from its centre was measured to be 2.7×10^{-3}.
- **a** Estimate the maximum speed of the star.
- **b** What assumption did you make in answering (a)?
- **c** Determine a value for the average density of the galaxy.

Fluctuations in the CMB

As we have seen, the discovery in the 1960s of isotropic microwave radiation (corresponding to a wavelength of 2.76 K) arriving at Earth from all directions was convincing evidence for the Big Bang model of the universe. This radiation is called cosmic microwave background radiation (CMB). In the last 25 years a great deal of research has been done into looking for *tiny* variations (fluctuations) in the CMB. Figure 16.57 shows an image of the fluctuations in CMB compiled from data from the Planck mission. Different temperatures are represented by different colours, but the maximum variation is only 0.0002 K! A significant 'cold spot' is ringed.

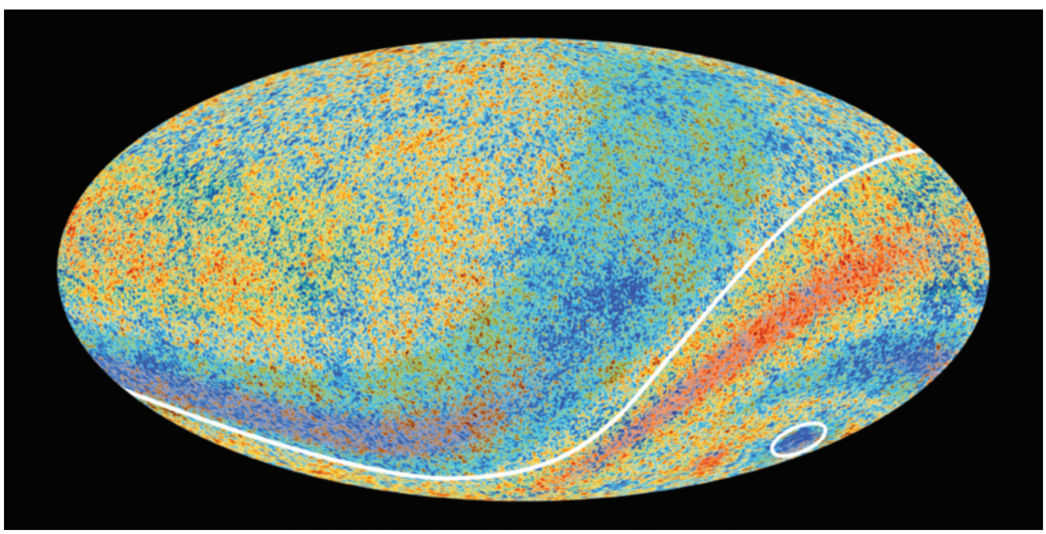


Figure 16.57 Fluctuations in the CMB as published by the Planck mission in 2013

It is clear from such images that the CMB is not *perfectly* isotropic. For example, there is a clear difference either side of the white line (which divides the opposite hemispheres of the sky). These variations in CMB are known as **anisotropies** and they provide astronomers with information about the early universe, at the time that the radiation was emitted. The early universe was opaque to electromagnetic radiation until it was 380 000 years old. This may seem like a long time, but it was very early in the history of the universe and knowledge from the CMB dates from that time. For example, the associated fluctuations in density may have been responsible for the later formation of galaxies and clusters of galaxies.

Linking average universe temperature to the cosmological scale factor

As the universe expands, the wavelength at which the maximum radiation intensity is detected, λ_{max}, also stretches and this is represented by the cosmic scale factor, R. That is:

$$\lambda_{max} \propto R$$

but Wien's law tells us $\lambda_{max}T$ = constant (2.9×10^{-3} m K), so that:

$$T \propto \frac{1}{R}$$

This equation is *not* given in the *Physics data booklet*.

Worked example

12 The average temperature of the universe now (when $R = 1$) is 2.76 K.

a What was the cosmic scale factor when the average temperature was 50 K?

b The distance between Earth and a certain galaxy is now 4.0×10^8 ly. What was the distance when the average temperature of the universe was 50 K?

a $T \propto \frac{1}{R}$ or TR = constant

$(TR)_{now} = (TR)_{then}$

$2.76 \times 1 = 50R_{then}$

$R_{then} = 0.056$

b $\frac{R_{now}}{R_{then}} = \frac{d_{now}}{d_{then}}$

$\frac{1}{0.056} = \frac{4.0 \times 10^8}{d_{then}}$

$d_{then} = 2.2 \times 10^7$ ly

CMB missions

The importance of the CMB is demonstrated by the fact that four major satellite programmes have involved it:

- **COBE** – Cosmic Background Explorer, launched in 1989
- **WMAP** – Wilkinson Microwave Anisotropy Probe, launched in 2001
- **Planck mission** – launched in 2009
- **James Webb telescope** – scheduled for launch in 2018.

The improving technology for successive missions has resulted in rapidly increasing precision and quantity of data. Comparison of Figure 16.43 (COBE and WMAP) with Figure 16.57 (Planck mission) illustrates the improved resolution.

The Planck mission has now ended, although its data is still being examined. In March 2013 the European Space Agency published Figure 16.57 and their latest estimates for some important astronomical data:

- The universe is 13.798 ± 0.037 billion years old.
- The universe contains 4.9% ordinary matter, 26.8% dark matter and 68.3% dark energy.
- The Hubble constant was measured to be 67.80 ± 0.77 km s^{-1} Mpc^{-1}. (This figure is significantly different from other recent quoted values, e.g. from WMAP.)

86 Explain why the average temperature of the universe is inversely proportional to the cosmic scale factor.

87 In the distant future the average temperature of the universe may cool to 1 K. What will the cosmic scale factor be at that time?

88 **a** Use the internet to find out what the latest accepted value is for the Hubble constant.

b Why is it sometimes called the Hubble *parameter*?

89 Find out more about the proposed James Webb telescope.

Summary of knowledge

16.1 Stellar quantities

- Nebulae are enormous diffuse 'clouds' of interstellar matter, mainly gases (mostly hydrogen and helium) and dust. Large nebulae are the principal locations for the formation of stars.
- Over a very long period of time, gravity pulls atoms closer together and eventually they can gain very high kinetic energies (that is, the temperature becomes extremely high – millions of kelvin) if the overall mass is large. The hydrogen nuclei (protons) can then have enough kinetic energy to overcome the very high electric forces of repulsion between them and fuse together to make helium. When this happens on a large scale it is called the birth of a (main sequence) star.
- A main sequence star can remain in equilibrium for a long time because the gravitational pressure inwards is balanced by thermal gas pressure and radiation pressure outwards.
- Many spots of light that seem to be stars are in fact binary stars, with two stars orbiting their common centre of mass.
- The forces of gravity cause billions of stars to collect in groups (galaxies), orbiting a common centre of mass. Galaxies also form into groups called clusters of galaxies. These clusters are not distributed randomly in space and are themselves grouped in super clusters (the largest structures in the universe).
- Stars formed from the same nebula within a galaxy may also form groups called stellar clusters, which are bound together by gravity and move together. Globular clusters contain large numbers of stars so that gravity forms them into roughly spherical shapes. Open clusters are newer and have fewer stars in less-well-defined shapes.
- Stellar clusters should not be confused with constellations, which are simply patterns of stars as seen from Earth. The stars in a constellation may have no connection to each other and may not even be relatively close, despite appearances.
- Planetary systems, like our solar system, are formed around some stars in the same process that created the star. The planets move in elliptical paths with periods that depend on the distance from the star. Comets are much smaller than planets, with typically much longer periods and more elliptical paths. When they are close to the Sun (and the Earth) they may become visible to us, and may have a 'tail' of particles created by the solar wind.
- Astronomers use several different units for measuring distance. The light year, ly, is defined as the distance travelled by light in a vacuum in 1 year. The astronomical unit (AU) is equal to the mean distance between the Earth and the Sun. The parsec (pc) = 3.26 ly.
- The order of magnitude of the diameter of the observable universe is 10^{11} ly. A typical galaxy has a diameter of about 10^4 ly and the distance between galaxies is typically 10^7 ly.
- The measurement of astronomical distances is a key issue in the study of astronomy. The distance to nearby stars can be calculated from a measurement of the parallax angle between their apparent positions (against a background of more distant, fixed stars) at two times separated by 6 months.
- One parsec is defined as the distance to a star that has a parallax angle of one arc-second.
- d (parsec) = $1/p$ (arc-seconds)
- For stars further away than a few hundred parsecs the stellar parallax method is not possible because the parallax angle is too small to measure accurately.
- The apparent brightness, b, of a star (including the Sun) is defined as the intensity (power/perpendicular receiving area) on Earth. The units are W m^{-2}.
- The luminosity, L, of a star is defined as the total power it radiates (in the form of electromagnetic waves). It is measured in watts, W.

- Understanding the relationship between luminosity and apparent brightness is very important in the study of astronomy – apparent brightness, $b = L/4\pi d^2$, where d is the distance between the star and Earth. This assumes that the radiation spreads equally in all directions without absorption in the intervening space. For very distant stars this assumption can lead to inaccuracies.
- The luminosity (power) of a star can be determined from the Stefan–Boltzmann law (Chapter 8): $P = e\sigma AT^4$, which reduces to $L = \sigma AT^4$ if we assume that stars behave like perfect black bodies and so have emissivities of 1.

16.2 Stellar characteristics and stellar evolution

- Intensity–wavelength graphs are very useful for representing and comparing the black-body radiation from stars with different surface temperatures. Such graphs can be used to explain why stars emit slightly different colours.
- Wien's displacement law (Chapter 8) can be used to calculate the surface temperature of a star if the wavelength at which the maximum intensity is received can be measured: $\lambda_{max}T = 2.9 \times 10^{-3}\,\text{m K}$.
- The elements present in the outer layers of a star can be identified from the absorption spectrum of light received from the star.
- Stars that are formed from greater masses will have stronger gravitational forces pulling them together. This will result in higher temperatures at their core and faster rates of nuclear fusion. More massive main sequence stars will have bigger sizes, higher surface temperatures, brighter luminosities and shorter lifetimes.
- For main sequence stars the approximate relationship between mass and luminosity is represented by the equation $L \propto M^{3.5}$.
- The Hertzsprung–Russell (HR) diagram is a common way of representing different stars on the same chart. The (logarithmic) axes of the diagram are luminosity and temperature (reversed). The sizes of different stars can be compared if lines of constant radius are included on the diagram.
- The majority of stars are located somewhere along a diagonal line from top-left to bottom-right of the HR diagram. This is called the main sequence. The only basic difference between these stars is their mass – which results in different luminosities and temperatures because of the different rates of fusion.
- Other types of stars, like red giants, white dwarfs, supergiants and Cepheid variables (on the instability strip) can be located in other parts of the HR diagram.
- The outer layers of Cepheid variable stars expand and contract regularly under the competing influences of gravity and thermal gas pressure. The period of the resulting changes in the observed apparent brightness is related to the star's luminosity and represented in a well-known relationship, so that a Cepheid's luminosity can be determined from its period. $b = L/4\pi d^2$ can then be used to determine the distance, d, to the star and therefore the galaxy in which it is situated.
- Inaccuracies in the data involved mean that these estimates of distance, especially to the furthest galaxies, are uncertain. This uncertainty has been a significant problem when estimating the age of the universe.
- When the supply of hydrogen in a main sequence star reduces below a certain value, the previous equilibrium is not sustained and the core will begin to collapse inwards. Gravitational energy is again transferred to kinetic energy of the particles and the temperature of the core rises even higher than before. This results in the outer layers of the star expanding considerably and, therefore, cooling. It is then possible for the helium in the core to fuse together to form carbon and possibly some larger nuclei, releasing more energy so that the star becomes more luminous. So, the star has a hotter core but it has become larger and cooler on the surface. Its colour therefore changes and it is then known as a red giant (or a red supergiant).

- After nuclear fusion in the core finishes, if the original mass of a red giant star was less than a certain value (about eight solar masses), the energy that is released as the core contracts forces the outer layers of the star to be ejected in what is known as a planetary nebula. The core of the star that is left behind has a much reduced mass. It is small and luminous and is described as a white dwarf. A white dwarf star can remain stable for a long time because of a process known as electron degeneracy pressure. The Chandrasekhar limit is the maximum mass of a white dwarf star (= 1.4 × solar mass).
- Red giants with original masses heavier than eight solar masses are known as red supergiants, but electron degeneracy pressure is not high enough to prevent further collapse and the resulting nuclear changes in the core produce a massive explosion called a supernova.
- If the core, after a supernova, has a mass of less than approximately three solar masses (called the Oppenheimer–Volkoff limit), it will contract to a very dense neutron star. It can remain stable for a long time because of a process known as neutron degeneracy pressure. If the mass is larger than the Oppenheimer–Volkoff limit, the core will collapse further to form a black hole.
- The changes to stars after they leave the main sequence can be traced on the HR diagram.

16.3 Cosmology

- When the line spectra emitted from galaxies are compared with the line spectra from the same elements emitted on Earth, the observed wavelengths (and frequencies) are slightly different. In most cases there is a very small increase (shift) in wavelengths, $\Delta\lambda$. Because red is at the higher wavelength end of the visible spectrum this change is commonly known as a 'red-shift'. More precisely, red-shift is defined by $z = \Delta\lambda/\lambda_0$, where λ_0 is the wavelength measured at source.
- Red-shift occurs because the distance between the galaxy and Earth is increasing. This is similar to the Doppler effect in which the wavelength of a source of sound that is moving away from us is increased.
- If the shift is to a longer wavelength (a red-shift), we know that the motion of the star or galaxy is away from Earth. We say that the star is receding from Earth.
- When the light from a large number of galaxies is studied, we find that nearly all the galaxies are receding from Earth, and each other. This can only mean that the universe is expanding.
- The magnitude of the red-shift, z (= $\Delta\lambda/\lambda_0$), can be shown to be approximately equal to the ratio of the recession speed to the speed of light $\approx v/c$. This equation can be used to determine the recession speed of galaxies, but it cannot be used for galaxies that are moving at speeds close to the speed of light.
- The light from a small number of stars and galaxies is blue-shifted because their rotational speed within their galaxy or cluster of galaxies is faster than the recession speed of the whole system.
- A graph of recession speed, v, against distance from Earth, d, shows that the recession speed of a galaxy is proportional to its distance away. This is important evidence for the Big Bang model of the universe – the universe began at one point at a particular time (13.8 billion years ago) and has been expanding ever since. This was the creation of everything, including both space and time.
- Hubble's law is $v = H_0 d$ where H_0 is known as the Hubble constant (the gradient of the graph). The current value of the Hubble constant is not precisely known because of uncertainties in measurements of v and d. This equation can be used to estimate the age of the universe ($T = 1/H_0$), although it would wrongly assume that the universe has always been expanding at the same rate.
- It is important to understand that space itself is expanding, rather than galaxies expanding into a pre-existing empty space. The universe has no centre and no visible edge.
- The discovery of cosmic microwave background (CMB) radiation coming (almost) equally from all directions (isotropic) confirmed the Hot Big Bang model of the universe. The radiation is characteristic of a temperature of 2.76 K, which is the predicted temperature

to which the universe would have cooled since its creation. Alternatively, the current wavelength of CMB can be considered as a consequence of the expansion of space (the wavelength emitted billions of years ago was much smaller).

- Astronomers use the cosmic scale factor, R, to represent the size of the universe – R (at a time t) = the separation of two galaxies at time t the separation of the same two galaxies at some other selected time (usually now, so that the value of R now is 1).
- Red-shift is related to the cosmic scale factor by $z = (R/R_0) - 1$, where R_0 was the value of R at the time the radiation was emitted.
- Possible futures of the universe depend on if, and how, the expansion will continue. Simplified graphs of size of the universe (or cosmic scale factor) against time can be used to show the basic possibilities. They also represent different possibilities for the previous rates of expansion.
- The luminosities of Type Ia supernovae are known to be (almost) all the same, so that their distances from Earth can be calculated. However, recent measurements of their associated red-shifts suggest that these very distant stars are further away than the Hubble law predicts. In other words, the universe is expanding quicker than previously believed – the universe is 'accelerating'. It had been assumed that the forces of gravity would reduce the rate of expansion of the universe.
- The concept of 'dark energy' existing in very low concentration throughout space has been proposed as a possible explanation for the increasing rate of expansion of the universe.

16.4 Stellar processes

- The inwards collapse of clouds of interstellar matter because of gravitational forces is opposed by the random motions of the particles, creating an outwards pressure. In order for star formation to begin, the total mass of the cloud has to be high enough to create sufficient inwards gravitational forces. For a given temperature, the minimum mass required is called the Jeans mass, M_J. The Jeans mass is large enough for the formation of many stars from the same cloud.
- The Jeans criterion is that the collapse of an interstellar cloud to form a star can only begin if its mass, M, is higher than M_J.
- The fusion of hydrogen to helium in main sequence stars is a three-stage process known as the proton–proton cycle. It involves the release of a large amount of energy in the form of the kinetic energy of the nuclei, gamma rays and neutrinos.
- Because helium is denser than hydrogen, it remains at the centre of the star where it was formed.
- $L \propto M^{3.5}$ shows that more massive stars are much more luminous. If we assume that the luminosity of a star is proportional to mass/lifetime, T, then the lifetime of a main sequence star is approximately represented by $T \propto 1/M^{2.5}$. Using this equation, if we know the mass of a star, we can compare its lifetime to that of our Sun (the mass and lifetime of which are well known).
- A typical main sequence star begins its life with about 75% hydrogen and its main sequence lifetime will end when about 12% of its total hydrogen has been fused into helium. It will then form a red giant or supergiant, as described earlier.
- The temperatures in the cores of red giants and supergiants are sufficient to cause the fusion of helium nuclei (or heavier elements). The creation of the nuclei of heavier elements by fusion is called nucleosynthesis.
- For stellar masses lower than $4M_\odot$ (red giants) the core temperature can reach 10^8 K and this is large enough for the nucleosynthesis of carbon and oxygen.
- For stellar masses between $4M_\odot$ and $8M_\odot$ (large red giants) the core temperature can exceed 10^9 K and this is large enough for the nucleosynthesis of neon and magnesium.

- For stellar masses heavier than $8M_{\odot}$ (red supergiants) the core temperature is high enough for the nucleosynthesis of elements as heavy as silicon and iron. Because iron and nickel are the most stable nuclei, the nucleosynthesis of heavier elements by fusion is not possible.
- The structure of stars off the main sequence is layered ('like the skins of an onion') as the heavier elements are formed closer to the centre. A red supergiant will have the most layers.
- The formation of elements heavier than iron involves the processes of repeated neutron capture. Many neutrons are released during fusion processes in stars. Because they are uncharged they are not repelled by other nucleons and they can enter nuclei and be 'captured' by strong nuclear forces.
- Slow neutron capture (s-process) can occur in red giants over long periods of time at intermediate temperatures and neutron densities. The new nucleus formed undergoes beta decay to an element with increased proton number.
- Rapid neutron capture (r-process) can occur very quickly in supernovae at extreme temperatures and neutron densities. Many neutrons are captured by the same nucleus before beta decay occurs. The heaviest elements are formed this way.
- Supernovae are sudden, unpredictable and very luminous stellar explosions.
- Type Ia supernovae occur when a white dwarf star attracts enough matter from a close neighbour (in a binary system) to increase its mass sufficiently so that electron degeneracy pressure is no longer sufficient to resist its sudden collapse and increase in temperature. Widespread and sudden fusion results in the explosion. As explained before, this process occurs at a precise mass and the resulting luminosity is always the same, such that they are used as 'standard candles' to determine the distance to remote galaxies.
- Type II supernovae are the result of the inwards collapse of red supergiants when the fusion processes stop.

16.5 Further cosmology

- The cosmological principle states that (on the large scale) the universe is homogeneous and isotropic.
- There are two main reasons why radiation from a star or galaxy may be red-shifted:
  - The expansion of the universe. The space between the source and the observer has expanded between the time the radiation was emitted and the time it was received. This is called the cosmological red-shift.
  - The source of radiation and the observer could be moving relative to each other in unchanging space. This is known as the Doppler effect and it can also result in a blue-shift if the source and observer are moving closer together.
- It is possible for cosmological red-shift and the Doppler shift to occur at the same time, for example if a star is moving in its galaxy towards Earth, while the galaxy as a whole is receding due to the expansion of space.
- To understand the origins and future of the universe we seem to need to know how much mass it contains. But astronomers refer to the average density of the universe because it is assumed to be homogeneous and we are unable to observe all of it.
- The critical density, ρ_c, of the universe is the theoretical density that would *just* stop the expansion of the universe after an infinite time.
- Critical density can be related to the Hubble constant using classical physics: $\rho_c = 3H^2/8\pi G$.
- The actual average density of the universe can be estimated from the masses of the galaxies and their distribution. But adding the masses of the observed stars in a galaxy together does not produce an accurate figure for its total mass.
- Classical physics theory can be used to plot a graph of the orbital speed of stars in a galaxy against their distance, r, from the centre (a 'rotation curve').

- For stars close to the centre, the theoretical equation:

$$v = \sqrt{\frac{4\pi G\rho}{3}}\,r$$

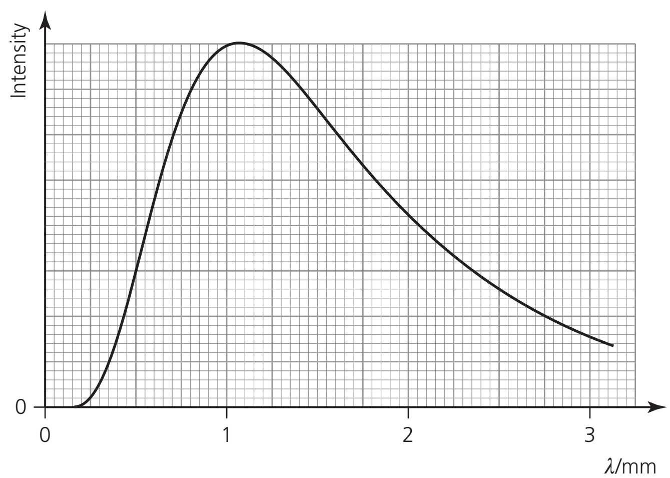
predicts the actual speeds reasonably accurately, but for more distant stars the prediction of rotational speeds does not match actual observations (using red-shift measurements). The more distant stars rotate at much higher speeds than expected. Astronomers explain this by proposing that the galaxy contains a large amount of matter that cannot be detected, called 'dark matter' (mostly in an outer halo).
- Dark matter is the name given to the proposed matter that must be present in the universe, but which has never been detected because it neither emits nor absorbs radiation.
- Dark matter is a subject of much continuing research in astronomy as explanations for this 'missing' mass are sought. MACHOs and WIMPs (including neutrinos) are two possible categories of particle that may explain dark matter.
- If the actual average density of the universe equals the critical density, the universe will expand at a decreasing rate, which will become zero after an infinite time – this is called a flat universe.
- If the actual average density of the universe is higher than the critical density, the universe will expand to a maximum size and then contract back to a point – this is called a closed universe.
- If the actual average density of the universe is lower than the critical density, the universe will continue to expand forever – this is called an open universe.
- The latest calculations (including dark matter) suggest that the actual average density of the universe is close to the critical density.
- However, the latest measurements on the red-shifts from distant supernovae provide strong evidence that the universe has been expanding at an increasing rate for about half of its lifetime. As mentioned before, this has led to the proposal of dark energy permeating all space.
- All of these possible futures for the universe can be represented on graphs of cosmic scale factor against time.
- From Wien's law we know that $\lambda_{max} \propto 1/T$ and as the age of the universe increases, space expands, λ_{max} increases and the cosmic scale factor R increases in proportion, so that $T \propto 1/R$. The average temperature of the universe multiplied by the cosmic scale factor is a constant.
- Fluctuations in the CMB (anisotropies) have been the focus of much research in recent years. Although these variations are tiny, they provide important evidence about the early stages of the universe and the origin of galaxies.
- The COBE, WMAP and Planck missions have provided an ever-improving bank of data on which astronomers are building an impressive understanding of the universe. Apart from information about anisotropies, this includes the latest estimates for the critical density and the age of the universe, plus estimates of the proportions of observable mass, dark matter and dark energy in the universe.

Examination questions – a selection

Paper 3 IB questions and IB style questions

Q1 **a** **i** What is the main energy source of a star? (1)
ii Explain how it is possible for a main sequence star to remain stable for billions of years. (2)
b **i** Define the luminosity of a star. (1)
ii Explain why main sequence stars can have very different luminosities. (2)
c **i** Define the apparent brightness of a star. (2)
ii Give two reasons why stars may have different apparent brightnesses. (2)
d Antares is a red supergiant 170 pc from Earth. Its luminosity is 2.5×10^{31} W and its surface temperature is 3400 K.
i Calculate the apparent brightness of Antares as seen from Earth. (2)
ii Explain what is meant by the term *red supergiant*. (2)
iii At what wavelength is the maximum intensity of the spectrum from Antares? (2)

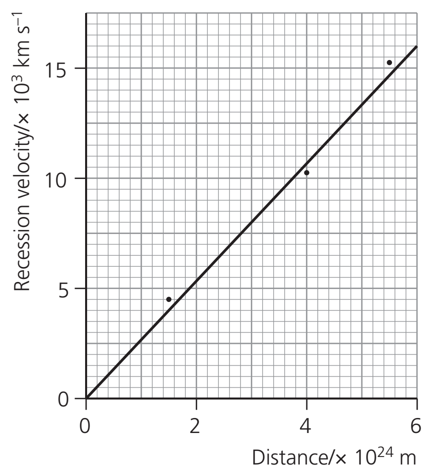
Q2 This question is about cosmic microwave background radiation. The graph shows the spectrum of the cosmic microwave background radiation.



The shape of the graph suggests a black-body spectrum, i.e. a spectrum to which the Wien displacement law applies.
a Use the graph to estimate the black-body temperature. (2)
b Explain how your answer to **a** is evidence in support of the Big Bang model. (2)
c State and explain another piece of experimental evidence in support of the Big Bang model. (2)

Q3 This question is about the mass–luminosity relation and also the evolution of stars.
The mass–luminosity relation for main sequence stars is assumed to be $L \propto M^{3.5}$, where L is the luminosity and M is the mass. Star X is 8×10^4 times more luminous than the Sun and 25 times more massive than the Sun.
a Deduce that star X is a main sequence star. (2)
b Outline with reference to the Oppenheimer–Volkoff limit, the evolutionary steps and the fate of star X after it leaves the main sequence. (3)

Q4 This question is about Hubble's law and the expansion of the universe.
a The spectrum of the cluster of galaxies Pegasus I shows a shift of 5.04 nm in the wavelength of the K-line. The wavelength of this line from a laboratory source is measured as 396.8 nm. Calculate the velocity of recession of the cluster. (2)
b The graph shows the recession velocities of a number of clusters of galaxies as a function of their approximate distances.

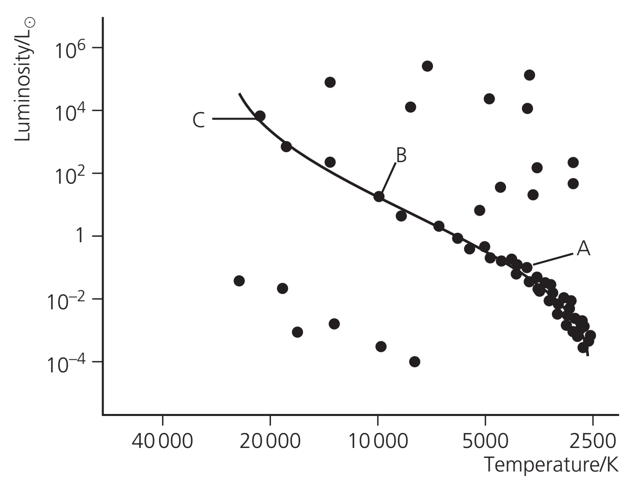


i State *one* method by which the distances shown on the graph could have been determined. (1)

ii Use the graph to show that the age of the universe is about 10^{17} s. (2)

© *IB Organization*

Q5 This question is about stellar evolution.
The diagram below represents a Hertzsprung-Russell (HR) diagram. The three identified stars (A, B and C) are all on the main sequence.



a Explain which of these stars is most likely to evolve to a white dwarf star.

b Draw and label the evolutionary path of the star as it evolves to a white dwarf star.

© *IB Organization*

Q6 a Explain how the cosmic scale factor is used by astronomers to represent the expansion of the universe. (2)

b If at some time, $+\Delta t$ in the future, the redshift of a distant galaxy is determined to be 0.020 calculate the cosmic scale factor at that time, compared with the present value of 1.00. (1)

c Suggest a possible value for the cosmic scale factor at a time $-\Delta t$ in the past. Explain your answer. (3)

Higher Level only

Q7 a What is a supernova? (1)

b Distinguish between the origins of type Ia supernovae and type II supernovae. (2)

c Explain why type Ia supernovae are considered to be ‘standard candles’. (3)

d Describe the *r process* in supernovae, which results in the creation of heavy elements. (2)

Q8 **a** Explain why stars can only be formed from nebulae of sufficient mass (the Jeans mass). (2)

b Explain why more massive stars have shorter lifetimes on the main sequence. (2)

c **i** If a star has a mass five times greater than the Sun, estimate its main sequence lifetime (compared with the Sun). (2)

ii What will happen to this star after its time on the main sequence? (2)

Q9 **a** Show that the speed of rotation of a star at a relatively close distance of 4 kpc from the centre of a rotating galaxy of average density $1 \times 10^{-20}\,\text{kg}\,\text{m}^{-3}$ is approximately $200\,\text{km}\,\text{s}^{-1}$. (2)

b Sketch a graph to show how the rotational speeds of stars within the galaxy vary with distance from the centre. (3)

c Explain how your graph indicates the existence of dark matter within the galaxy. (2)

d State one possible origin of dark matter. (1)

Q10 **a** Explain what is meant by the Cosmological Principle. (2)

b Observation of the night sky indicates that more stars can be seen in some directions than in other directions. Discuss whether this contradicts the Cosmological Principle. (2)

Q11 **a** Explain the concept of the critical density of the universe. (2)

b Determine a value for the critical density of the universe in $\text{kg}\,\text{m}^{-3}$ if the Hubble constant has a value of $70\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1}$. (2)

c **i** Express the critical density in terms of nucleons per cubic metre. (1)

ii Use your answer to estimate an average distance between nucleons at the critical density. (1)

d The actual average density of the universe is believed to be close in value to the critical density. However, the expansion of the universe is believed to be accelerating. Outline:

i the experimental evidence for this accelerating expansion (2)

ii how the proposed existence of dark energy may explain the expansion. (2)

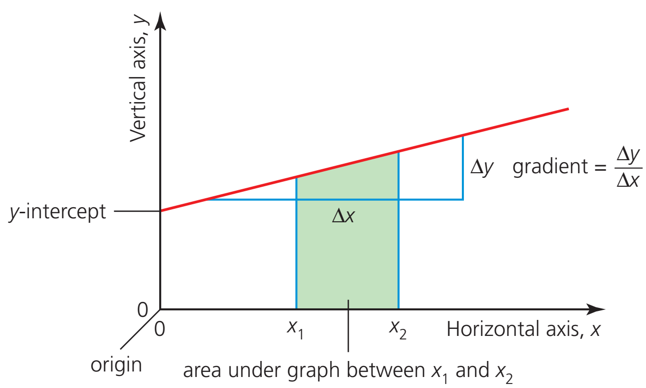
17 Appendix 1: Graphs and data analysis

Representing data graphically

There are many quantities that can be measured in a physics experiment. Usually, all but two of them are **controlled** so that they do not change during the course of the experiment. Then one quantity (called the **independent variable**) is deliberately varied or changed and the effect on one other quantity (called the **dependent variable**) investigated.

The best way to analyse the results of such experiments is often to **plot** (draw) a graph. Looking at a graph is a good way to identify a pattern, or trend, in numerical data. Graphs can also provide extra information – **gradients** (slopes), **intercepts** and the **areas** under graphs often have important meanings. Figure 17.1 illustrates the terminology associated with graphs.

Figure 17.1 Terminology for graphs



Drawing graphs

The drawing of good-quality graphs is a very important skill in physics. The following points should be remembered when drawing a graph.

- The larger a graph is, the more precisely the points can be plotted. A simple rule is that the graph should occupy at least half the available space (in each direction).
- Each **axis** should be labelled with the quantity and the unit used (e.g. force/N, speed/m s^{-1}). Lists of results in tables should be similarly labelled. For example, if you want to record a mass of 5 g as the number five on the axes of a graph, then labelling the axes as mass/g indicates that you have divided 5 g by g to get five.
- The *independent variable* is usually plotted on the horizontal axis and the *dependent variable* on the vertical axis. Sometimes, the choice of what to plot on each axis is made so that the gradient of the graph has a particular meaning. If time is one of the variables, it is nearly always plotted on the horizontal axis.
- The **scales** chosen should make plotting the points and interpreting the graph easy. For example, five divisions might be used to represent 10 or 20, but not 7 or 12.
- Usually, both scales should start at zero, so that the point (0, 0), the **origin**, is included on the graph. This is often important when interpreting a graph. However, this is not always sensible, especially if it would mean that all the readings were restricted to a small part of the graph. Temperature scales in °C do not usually need to start at zero.
- Data **points** should be neat and small. If points are used (rather that crosses), drawing a small circle around them can make sure that they are not overlooked, especially if the line goes through them.
- The more points that can be plotted, the more precisely the line representing the relationship can be drawn. At *least* six points are usually needed, although this may not always be possible.

- When all the points have been plotted, a pattern will usually be clear and a **line of best fit** can be drawn (see Figure 17.2 for two correct examples and one wrong example, in which the points are represented by error bars). These lines are sometimes called *trend lines*. Trend lines may be straight (drawn with a ruler) or a curve. (A straight line is described as being **linear**.) The line should be smooth and thin. Bumpy lines that try to pass through, or near, all the points show that the person drawing the line did not understand that points cannot be perfectly placed, and that there is uncertainty in all measurements. Typically there will be about as many points above a line of best fit as there are below the line. Points on a graph should *never* be joined by a series of straight lines.
- Drawing graphs by hand is a skill that all students should practise. However, knowing how to use a computer program to plot graphs is also a very valuable and time-saving skill (especially for investigation work). A graph generated by a computer (or graphic display calculator) must be judged by the same standards as a hand-drawn graph and sometimes their best-fit lines are not well placed.

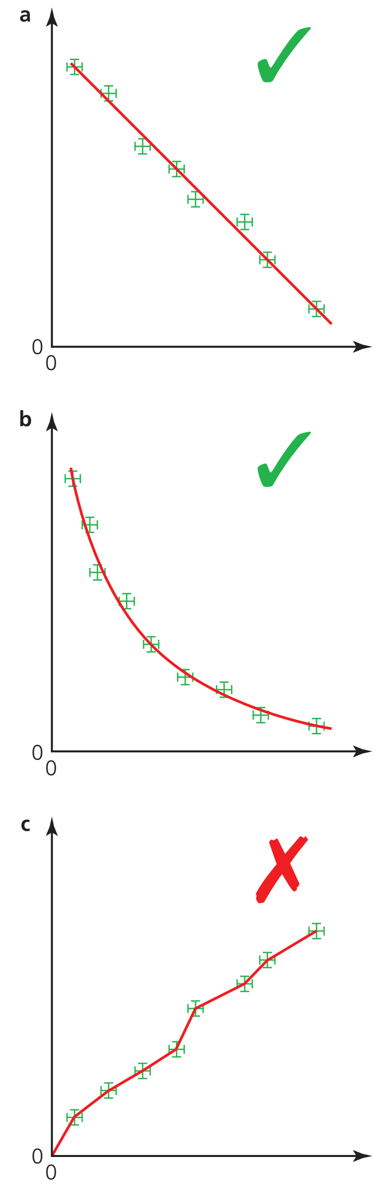


Figure 17.2 Right and wrong ways to draw best-fit lines

Extrapolating and interpolating

A line of best fit is usually drawn to cover a specific *range* of measurements recorded in an experiment, as shown in Figure 17.3. If we want to predict other values *within* that range, we can do that with confidence. The diagram indicates how a value for y can be determined for a chosen value of x. This is called **interpolation**.

If we want to predict what would happen *outside* the range of measurements (**extrapolation**) we need to *extend* the line of best fit. Lines are often extrapolated to see if they pass through the origin, or to find an intercept, as shown in Figure 17.3.

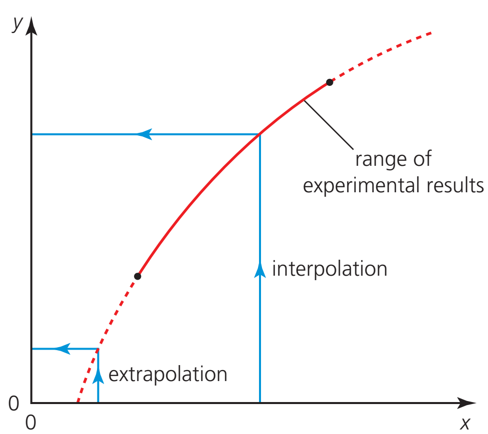


Figure 17.3 Interpolating and extrapolating to find the intercept on the *y*-axis

Predictions made by extrapolation should be treated with care, because it may be wrong to assume that the behaviour seen within the range of measurements also applies outside that range.

Proportionality

The simplest possible relationship between two variables is that they are **proportional** to each other (sometimes called *directly* proportional). This means that if one variable, say x, doubles, then the other variable, y, also doubles; if y is divided by five, then x is divided by five; if x is multiplied by 17, then y is multiplied by 17 etc. In other words, the ratio of the two variables (x/y or y/x) is constant. Proportionality is shown as:

$$y \propto x$$

Many basic experiments are aimed at investigating if there is a proportional relationship between two variables, and this is usually best checked by drawing a graph.

If two variables are (directly) proportional, then their graph will be a **straight line passing through the origin** (Figure 17.4). It is important to stress that a linear graph that does not pass through the origin *cannot* represent proportionality (Figure 17.5).

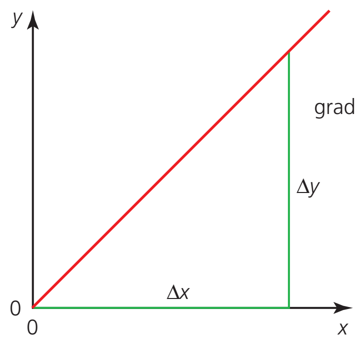


Figure 17.4 A proportional relationship

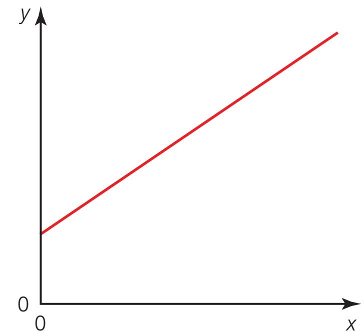


Figure 17.5 A linear relationship that is not proportional does not pass through the origin

Gradients of lines

The gradient of a line is given the symbol *m* and it is calculated by dividing a change in *y*, Δ*y*, by the corresponding change in *x*, Δ*x*, as shown in Figure 17.4. (A delta sign, Δ, is used to represent a change of a quantity.)

$$m = \frac{\Delta y}{\Delta x}$$

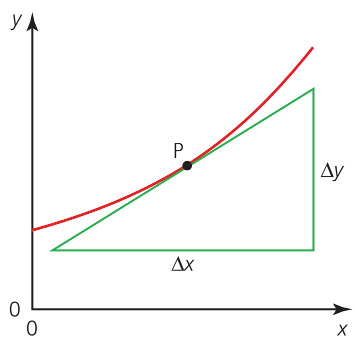


Figure 17.6 Finding the gradient of a curve at point P with a tangent

It is important to note that a *large* triangle should be used when determining the gradient of a line, because the percentage uncertainty will be less when using larger values.

The gradients of many lines have a physical meaning – for example the gradient of a mass – volume graph equals the density of the material.

The gradient of a curved line, such as that shown in Figure 17.6, is constantly changing. The gradient of the curve at any point, P, can be determined from a **tangent** drawn to the curve at that point. (In mathematics, if the equation of a line is known, then the gradient at any point can be determined by a process called *differentiation*.)

Worked examples

1 Figure 17.7 shows a best-fit line produced from an experiment in which the masses and volumes of different pieces of the same metal alloy were measured.

a Calculate a value for the density of the alloy, which is equal to the gradient of the line.

b Suggest why the graph does not pass through the origin.

c Explain why using the gradient to find the density is a much better method than just calculating a value from one pair of readings of mass and volume.

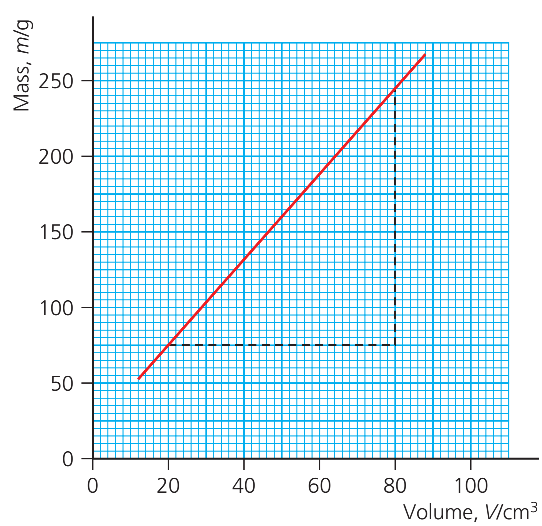


Figure 17.7 Graph of mass *m* against volume *V* of different pieces of alloy

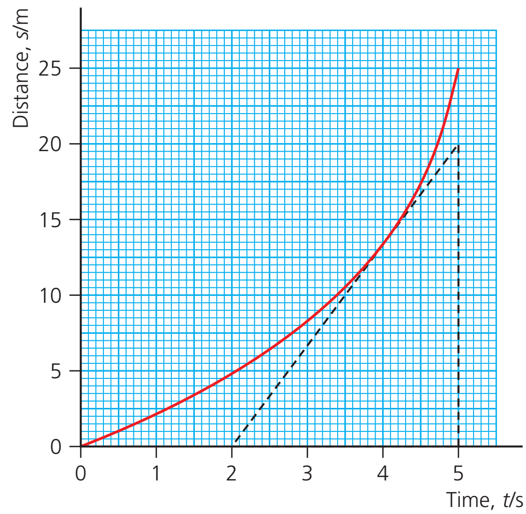
a Using the triangle shown on the graph:

$$\text{gradient} = \text{density} = \frac{\Delta m}{\Delta V} = \frac{245 - 75}{80 - 20} = 2.8\,\text{g}\,\text{cm}^{-3}$$

b The instrument used to measure mass had a zero offset error (of about +20 g).

c Individual readings may be inaccurate. The best-fit line reduces the effect of random errors and the zero offset error does not affect the result of the calculation.

2 Figure 17.8 shows a distance – time graph for an accelerating car. Determine the speed of the car after 4 s (equal to the gradient of the line at that time).



■ **Figure 17.8** Graph of distance *s* against time *t* for an accelerating car

The sloping dashed line is a tangent to the curve at $t = 4$ s.

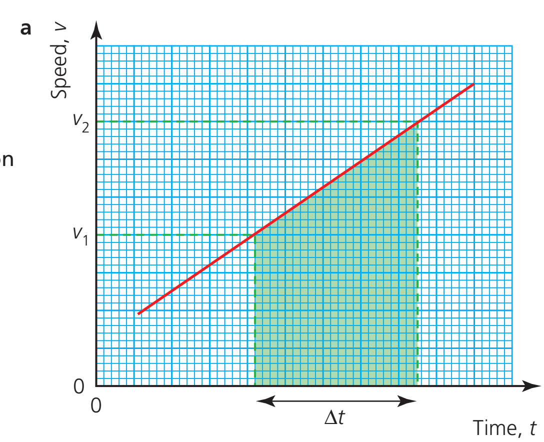
$$\text{gradient} = \text{speed} = \frac{\Delta s}{\Delta t} = \frac{20 - 0}{5.0 - 2.0} = \frac{20}{3.0} = 6.7\,\text{m}\,\text{s}^{-1}$$

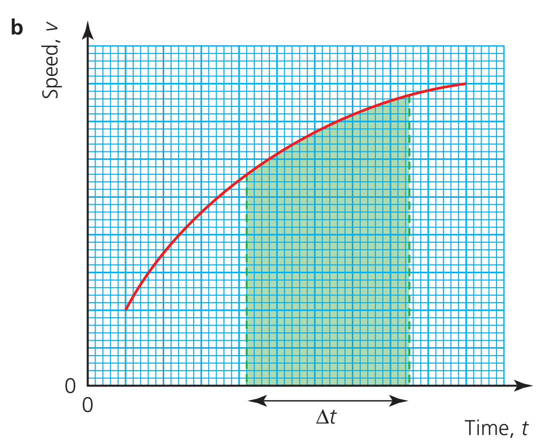
Areas under graphs

The area under many graphs has a physical meaning. As an example, consider Figure 17.9a, which shows part of a speed – time graph for a vehicle moving with constant acceleration. The area under the graph (the shaded area) can be calculated from the average speed, given by $\frac{(v_1 + v_2)}{2}$, multiplied by the time, Δt. The area under the graph is therefore equal to the distance travelled in time Δt.

In Figure 17.9b a vehicle is moving with a changing (decreasing) acceleration, so that the graph is curved, but the same rule applies – the area under the graph (shaded) represents the distance travelled in time Δt.

■ **Figure 17.9** Area under a speed–time graph for **a** constant acceleration and **b** changing acceleration

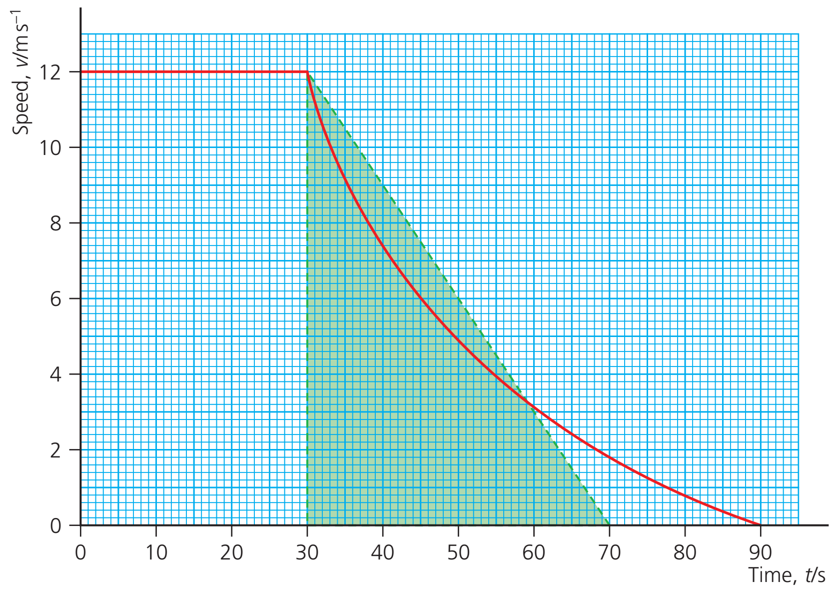




The area in Figure 17.9b can be estimated in a number of different ways, for example by counting small squares, or by drawing a rectangle that appears (as judged by eye) to have the same area. (If the equation of the line is known, it can be calculated using the process of *integration*.)

Worked example

3 Figure 17.10 represents the motion of a train that travels at a constant speed for 30 s and then decelerates for 60 s. Calculate the distance travelled in 90 s (equal to the area under the graph).



■ **Figure 17.10** Graph of velocity, *v*, against time, *t*, for a train

The area under the graph up to a time of 30 s = 12 × 30 = 360 m.

The area under the graph between 30 s and 90 s can be estimated from the shaded triangle, which has been drawn so that its area appears to be the same as the area under the curved line:

area = $\frac{1}{2}$ × 12 × (70 − 30) = 240 m

total distance (area) = 360 + 240 = 600 m

The usefulness of straight-line graphs

Straight lines are much easier to understand and analyse than curved lines, but when directly measured experimental data are plotted against each other (x and y, for example), the lines are often curved rather than linear.

Data that give an $x - y$ curve can be used to draw other graphs to check different possible relationships. For example:

- A graph of y against x^2 could be drawn to see if a straight line through the origin is obtained, which would confirm that y was proportional to x^2.
- A graph of y against $\frac{1}{x}$ that passed through the origin would confirm that y was proportional to $\frac{1}{x}$ (in which case x and y are said to be **inversely proportional** to each other).
- A graph of y against $\frac{1}{x^2}$ passing through the origin would represent an **inverse square relationship**.

Figure 17.11 shows graphs of the most common relationships.

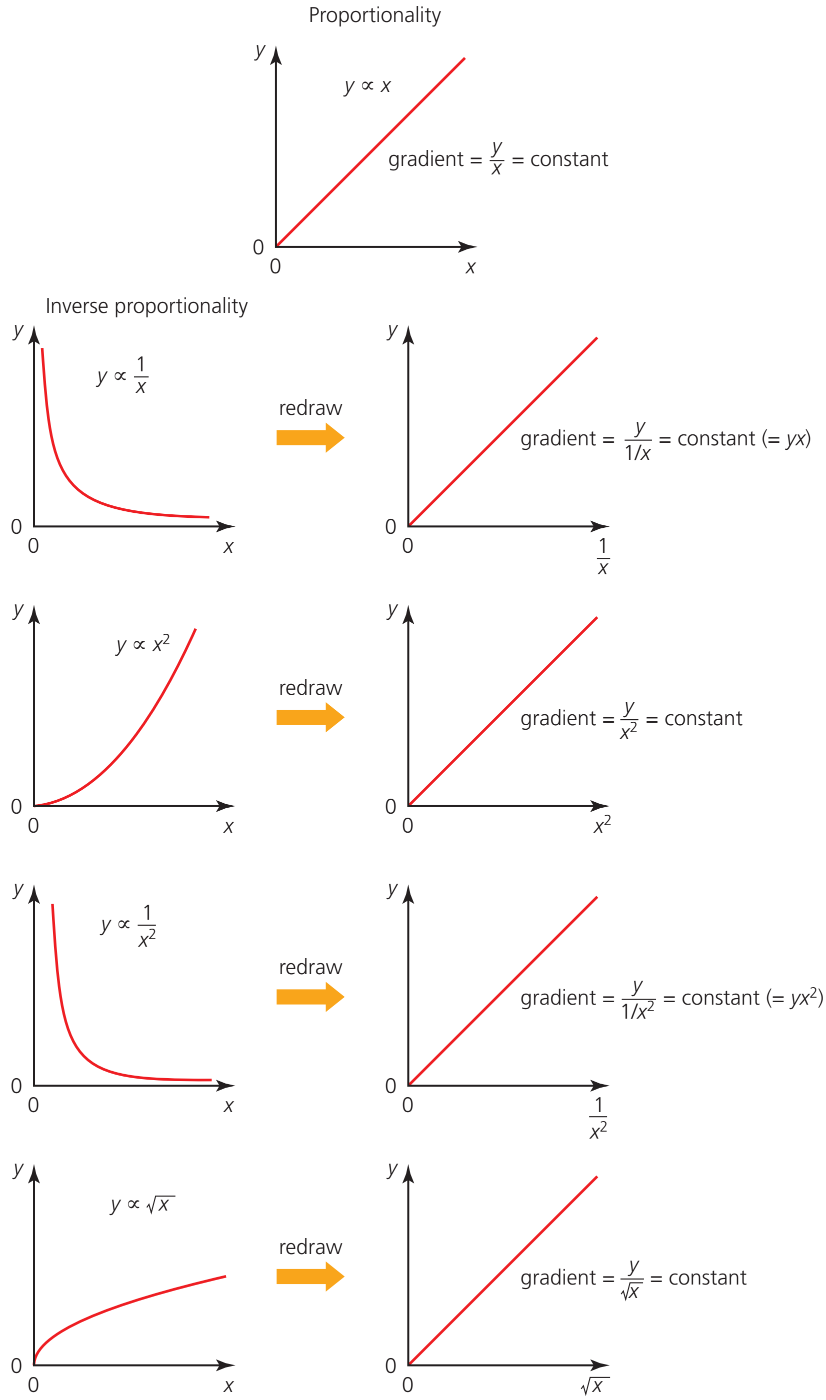


Figure 17.11 Some common graphical relationships, showing how curves can be replotted to produce straight lines

Worked example

4 In an internal assessment students were asked to investigate one factor that affected the deflection of a wooden beam fixed only at one end (a cantilever), and which had a mass (load) hanging from somewhere on the part that extended from the bench top (Figure 17.12).

A student listed the following variables: (i) type of wood, (ii) thickness of wood, (iii) width of wood, (iv) length of wood from where it was fixed, *L*, (v) position of load, (vi) mass of load. He decided to investigate how the deflection, *D*, depended on the length, *L*, keeping all the other variables constant. His results are shown in Table 17.1 (for simplicity, uncertainties have not been included).

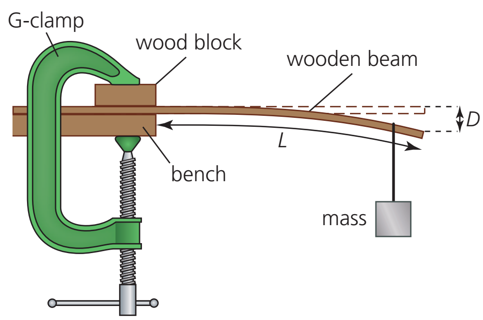


Figure 17.12

Table 17.1

L/cm	*D*/cm
30.0	0.8
40.0	2.0
50.0	3.7
60.0	6.5
70.0	10.8
80.0	15.8
90.0	22.2
100.0	31.2

a A graph of the raw data produces a curved line, so the student thought that maybe the deflection was proportional to the length squared or the length cubed. Perform numerical checks on the data to see if either of these possibilities is correct.

b Plot a suitable graph to confirm the correct relationship.

a If the deflection, *D*, is proportional to the length, *L* squared, $\frac{L^2}{D}$ (or $\frac{D}{L^2}$) will be constant, within the limits of experimental uncertainties. Calculations produce the following results (all $\times 10^2$ cm): 11.0, 8.0, 6.8, 5.5, 4.5, 4.1, 3.6, 3.2. These values are getting smaller for longer lengths, and are clearly *not* constant.

If the deflection, *D*, is proportional to the length, *L* cubed, $\frac{L^3}{D}$ (or $\frac{D}{L^3}$) will be constant, within the limits of experimental uncertainties. Calculations produce the following results (all $\times 10^4$ cm^2): 3.4, 3.2, 3.4, 3.3, 3.2, 3.2, 3.3, 3.2. These values are all very similar (within 3% of their average), confirming that the deflection is proportional to the length cubed.

b See Figure 17.13. A graph of *D* against L^3 produces a straight line through the origin. Note that it would have been better if the student had used lengths such that the points were spread out evenly along the line.

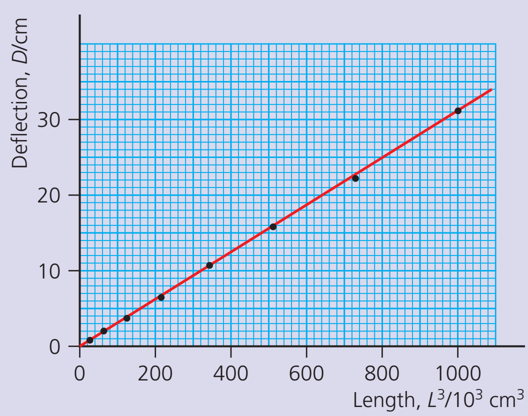


Figure 17.13

Equation of a straight line

All linear graphs can be represented by an equation of the form:

$$y = mx + c$$

where m is the gradient and c is the value of y when $x = 0$, known as the **y-intercept** (Figure 17.14).

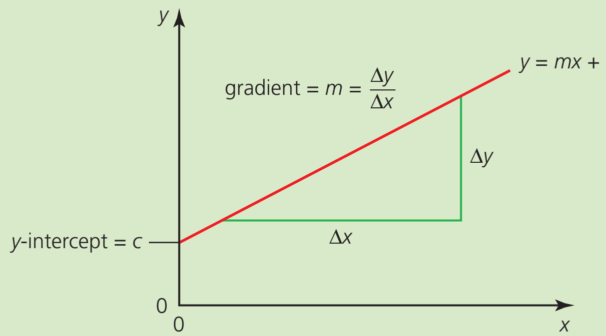


Figure 17.14 Graph of $y = mx + c$

Once a linear graph has been drawn, values for the gradient and the y-intercept can be determined and the results used to produce a mathematical equation to describe the relationship.

Worked example

5 Experimental data connecting two variables, x and y, are represented by the graph in Figure 17.15. Take measurements from the graph to enable you to write an equation to represent the relationship.

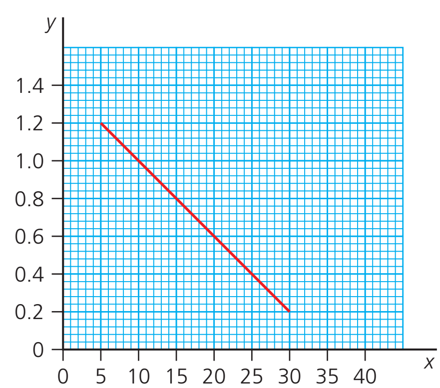


Figure 17.15

The gradient of the line, m, is $\frac{0.2 - 1.2}{30 - 5} = -0.04$ and the y-intercept, c, is 1.4.

Substituting into $y = mx + c$ we get:

$$y = -0.04x + 1.4$$

(which could be rewritten as $25y = 35 - x$).

Power laws and logarithmic graphs

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 constants. That is, y is proportional to x to the power p.

Taking logarithms of this equation we get:

$$\log y = p \log x + \log k$$

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 (Higher Level only).

Worked examples

6 The relationship between two variables, *x* and *y*, is shown in Figure 17.16. Take measurements from the graph so that you can write an equation to represent the relationship.

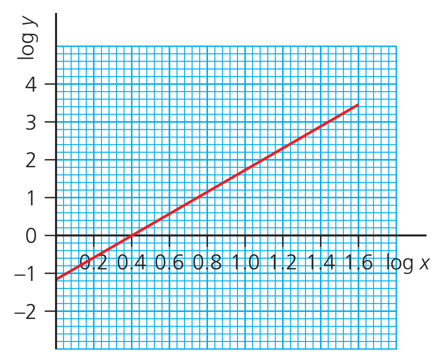


Figure 17.16

The gradient of the line, p, is 2.9 and the intercept on the $\log y$ axis, $\log k$, is –1.2. Substituting into $\log y = p \log x + \log k$ we get:

$$\log y = 2.9 \times \log x - 1.2$$

So,

$$y = 0.063x^{2.9}$$

7 Refer back to Worked example 4. Use the data to draw a log graph that verifies that the relationship is described by the equation $D = kL^m$. Determine values for m and k from the graph.

Taking logs of the equation:

$$\log D = m \log L + \log k$$

Comparing this to $y = mx + c$, we know that a graph of $\log D$ against $\log L$ will have a gradient of m, and k can be determined from an intercept (Figure 17.17).

$$m = 3.1 \text{ and } k = 2.2 \times 10^{-5}$$

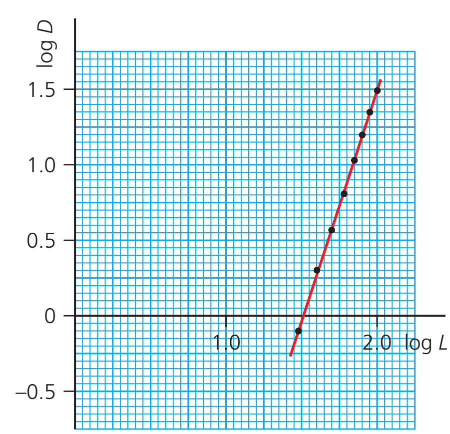


Figure 17.17

18 Appendix 2: Preparing for the IB Diploma Physics examination

Revision techniques

There are many different ways of revising for examinations, but some general words of advice apply to everyone:

- Use the IB syllabus and the summaries of knowledge for each chapter to identify your strengths and weaknesses at an early stage of your revision. Effective revision will concentrate on the parts of the syllabus with which you are less confident, rather than repeating topics you already know well. Monitor your progress, and you will be motivated by the fact that the list of topics left to revise becomes shorter and shorter.
- Revision should usually be *active* rather than passive. Discussing physics or answering questions is generally much more useful than reading or watching a physics video. However, it is sensible to *start* your revision of a topic by re-reading the appropriate section of this book.
- Answering questions from past examination papers is very important and many students and teachers believe that it is the best way to revise. You should do as many questions as you can, and check your answers, or have them checked by someone else. Completing 'mock exams', in which you answer all the required questions on complete examination papers in the regulation times, will also help you to judge whether you are working too quickly or slowly. You may also have taken a series of tests and examinations during your course; these are valuable resources for revision – we should all learn from our mistakes.
- Very few students enjoy revision so it is a good idea to use a variety of different revision techniques to stimulate interest. Some students also find that working in different surroundings can be a way to freshen up their revision. It is not a good idea to force yourself to revise when you are tired, nor to work for too long at one time. (Between 40 and 60 minutes may be the ideal length of time for revision without a break.)
- Most students find that planning a revision schedule helps to organize and structure their work. But it can be a waste of time if you don't stick to your schedule, so you need to keep re-writing it!
- You must be familiar with the structure of the examination papers and the styles of different exam questions (see below).

Examination paper details

The Physics data book is provided in all examinations. There is no choice of questions in any paper.

Make sure you take into the examination room all the equipment you may need: 30 cm ruler, protractor, compass, pens and pencils and a calculator of an approved type (not allowed in Paper 1).

	Standard Level	Higher Level
Paper 1	45 minutes	1 hour
	30 multiple-choice questions	45 multiple-choice questions
	Questions will be set on the Standard Level Core (Chapters 1–8)	Questions will be set on the Standard Level Core (Chapters 1–8) + the Additional Higher Level (AHL) content from Chapters 9–12
	Calculators are *not* allowed	Calculators are *not* allowed
	20% of total examination mark	20% of total examination mark
Paper 2	1 hour 15 minutes	2 hours 15 minutes
	Questions will be set on the Standard Level Core (Chapters 1–8)	Questions will be set on the Standard Level Core (Chapters 1–8) + AHL content from Chapters 9–12
	Short-answer and extended-response type questions	Short-answer and extended-response type questions
	Calculators are allowed	Calculators are allowed
	40% of total examination mark	36% of total examination mark
Paper 3	1 hour	1 hour 15 minutes
	Section A has one data-based type question and several short-answer questions on experimental work (Knowledge of Core material will be assumed (Chapters 1–8))	Section A has one data-based type question and several short-answer questions on experimental work. (Knowledge of Core and AHL material will be assumed (Chapters 1–12))
	Section B has short-answer and extended-response type questions on the Options (you select the questions for one Option)	Section B has short-answer and extended-response type questions on the Options (you select the questions for one Option)
	Calculators are allowed	Calculators are allowed
	20% of total examination mark	24% of total examination mark
Internal assessment	20% of total examination mark	

Taking the examination

There are very few students who (with the same knowledge of physics) could not improve their marks simply by improving their examination technique. Here are some tips.

Paper 1

- The questions are generally arranged in approximate syllabus order.
- Although multiple-choice questions are often considered to be easier than many of the questions in the other two papers, it is common for students to make careless mistakes. If you have time, double-check your answers.
- Never select any answer until you have read *all* the alternative possibilities.
- If you are unsure of an answer, do not spend too much time on it. Write comments next to the question, cross out any answers that seem obviously wrong (there are usually one or two!) and move on to the next question. If you work quickly enough you should have time to return to unfinished questions later – the question may seem easier second time around.
- Be aware that sometimes a possible answer is a correct statement in itself, but not the correct answer to the question.
- Look out for the inclusion of two answers that contradict each other. It is likely that one of them is correct (and the other wrong).
- Remember that there is no penalty for wrong answers. Never leave a question without an answer, even if it is only a guess.

Papers 2 and 3

- Read through the *whole* question before you begin to answer any part.
- Judge the amount of detail you need to supply in your answers from the amount of space allowed and the number of marks allocated to the question. In general you need to make a separate point for each mark.
- Read the questions very carefully and note or underline key words and phrases. If a question asks you to 'use' certain information, graphs or laws make sure that your answer does exactly that. If you answer the question in some other way, even correctly, you will not get the marks. If the question refers to a law or definition, begin your answer by quoting it.
- The presentation of your answers is important. Although neatness and correct spelling are not directly assessed, they create a favourable impression. Use a ruler to draw straight lines.
- Always show all your working in calculations.
- Give your answers in decimals not fractions. Use scientific notation wherever appropriate.
- Use appropriate significant figures in your answers and do not forget units.
- In general, it is better to express physical quantities in words, rather than symbols, although standard symbols can be used in the working of calculations.
- If you need to change one of yours answers, neatly cross out the work you want to delete. If there is not enough space for your new work, use extra pages and attach them to your answer booklet at the end of the examination.

Examples of different styles of exam question (Papers 2 and 3)

Command terms

All Paper 2 and 3 questions in IB Diploma Physics examinations contain one of a limited number of clear instructions, such as *define*, *outline* and *calculate*. These one-word instructions are known as *command terms* and they indicate the way in which you should answer the question. The command terms can be divided into three groups:

- demonstrate understanding (AO1)
- apply and use (AO2)
- construct, analyse and evaluate (AO3).

Demonstrate understanding

These command terms will be used to test your memory of factual knowledge of the syllabus.

Define

You are required to give the *exact* meaning of a word, phrase or quantity. You are strongly advised to make sure that you know all these definitions before the examination.
A surprisingly high proportion of students fail to achieve these 'easy' marks.

A definition should usually be written in words, rather than as an equation, although an equation is acceptable if the meanings of all the symbols are explained.

Worked example

Question: *Define* resistance.

Answer: Resistance is the ratio of the potential difference across a conductor to the current passing through it. ($R = \frac{V}{I}$ is only acceptable if the symbols are explained.)

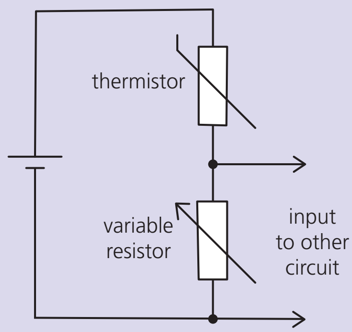
Draw

This widely used command term usually requires you to add something to an existing diagram or graph, but sometimes you may be asked to draw a completely new diagram. Use a sharp pencil and a ruler for drawing. Drawing a line of best fit on a graph is a common question (see Chapter 17).

Worked example

Question: *Draw* a circuit to show how a cell, a thermistor and a variable resistor can be connected to provide a potential dividing input to another circuit.

Answer:



Note that, in this example, it is not essential to label the components if standard electrical circuit symbols are used (see the *Physics data booklet*). The question is partly aimed at testing whether or not you know these symbols. If you are unsure of the correct symbol, draw a box and write beside it what it is meant to represent.

Label

This common instruction is often combined with an instruction to draw something. It may refer to an existing diagram, or an addition to a diagram that you are asked to make, or occasionally to a new drawing. The labels should normally be words rather than symbols. It is important to do this neatly and clearly, so the examiner can be sure exactly what you are labelling.

Worked example

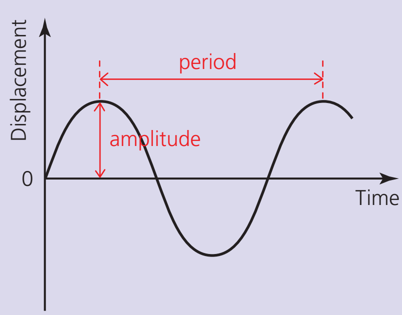
Question: Draw an arrow on the diagram to show the direction of the current and *label* it with the letter *I*.

This could be testing if you know that current flows conventionally from positive to negative around a circuit. The letter *I* should be placed close to the arrow. Without the labelling you may not be awarded the mark, even if the direction is correct.

Worked example

Question: Label the drawing to show the meaning of amplitude and period.

Answer:



Worked example

Question: The diagram shows a car moving at constant velocity. Draw fully *labelled* vectors to indicate the forces acting on the car.

Answer: At least four force vectors should be labelled as: weight, normal contact force, air resistance and push of road on tyres. The pairs of vectors should have equal lengths.

List

This means that you should provide a series of items *without* explanation. It should be clear from the question if the terms need to be in any order.

Worked example

Question: *List* the main energy sources used around the world to generate electrical power.

Answer: coal, nuclear, oil, natural gas, hydroelectric. (In this answer the order of the list is not important, nor is the list complete. A complete list would not be required by the examiner because the question is not intended to be definitive.)

Measure

This command term requires that you measure the value of a physical quantity from a diagram or graph on the examination paper. It could be a length, an area, an angle, or may require interpretation from a scale. Clearly you should be prepared by taking a 30 cm ruler and a protractor into the examination room. Obviously, measurement has to be accurate in order to get the marks available.

Worked example

Question: The diagram shows a ray of light being refracted as it enters some glass. Take *measurements* in order to calculate the refractive index of the medium. (This combines measurements with a calculation.)

Answer: It will be necessary to measure the angles of incidence and refraction.

State

State is similar to *define* in that a short, precise answer is required, without the need for any further explanation. This is one of the most widely used command terms in IB Diploma physics examinations. The syllabus contains some important laws and terms that may need to be 'stated' in an examination, and these should be memorized in the same way as definitions.

Worked example

Question: *State* what is meant by damping.

Answer: Damping is the dissipation of the energy of an oscillation when it is acted on by a resistive force.

Write down

This requires only a short, straightforward answer without explanation. The information is often readily available from within the question.

Worked example

Question: *Write down* the name of the source of most radiant energy arriving at the Earth.

Answer: The Sun.

Apply and use

These command terms will test your ability to *use* the concepts and principles of physics that you have learned during the course.

Annotate and apply

These are unusual command terms in IB Diploma Physics examinations.

- To *annotate* is similar to *label*, but requires *brief notes* to be added to a diagram or graph.
- To *apply* means to use knowledge in a new situation. For example, you could be asked to *apply* your understanding of Newton's laws of motion to a fairground ride.

Calculate

In this very common type of question you are required to use data given in the question and/or in the *Physics data booklet* in order to determine a numerical answer.

- You must show clearly *how* you obtained your answer. Marks are usually awarded for correct working, even if your final answer is wrong.
- Your answer must have a suitable number of significant figures (see Chapter 1).
- Your answer must have a unit (unless it is a ratio).

Worked example

Question: A stone is thrown vertically upwards with a speed of $8.10\,\mathrm{m\,s^{-1}}$. Ignoring air resistance, *calculate* the maximum height reached by the stone.

Answer:

$v^2 = u^2 + 2as$ (Quote the equation you are using.)

$0 = 8.10^2 + (2 \times -9.81 \times s)$ (Substitute the data. The value for a was taken from the *Physics data booklet*.)

$s = 3.34\,\mathrm{m}$

Answers of 3.3 m or 3.344 m will also be accepted.

Describe

Use knowledge from the course and/or information given in the question to give a straightforward account. The amount of detail needed will vary from question to question, and is best judged from the number of marks available for the answer.

The command term *describe* appears frequently in the IB examinations.

Worked example

Question: *Describe* what is meant by the term resonance. (3 marks)

Answer: Resonance is the increase in the amplitude of an oscillation when it is disturbed by an external force that has the same frequency as the natural frequency of the oscillator.

Note that there are three separate ideas included in this answer – in response to the 3 marks allocated to the question.

Distinguish

This command term means you should explain the essential difference(s) between two (or more) things. You may also briefly indicate what they have in common.

Worked example

Question: *Distinguish* between speed and velocity.

Answer: Speed is calculated from distance/time. Velocity has the same magnitude as speed, but a direction of motion must also be given. (Examples might help, but are not essential, unless asked for in the question.)

Estimate

This is similar to *calculate*, but an accurate answer will not be possible. For example, the question may involve you making a calculation based on *your* reasonable estimates of unknown quantities. Estimated answers should be given with an appropriate number of significant figures (which may be only one), or just an order of magnitude.

Making estimates is demanding for many students, but marks will be awarded for *any* reasonable estimates, rather than an expected answer. Your assumptions should be stated clearly.

Worked example

Question: *Estimate* the amount of coal that would be burned in a 100 MW power station in one hour.

Answer:

$$\text{mass of coal needed every second} = \frac{\text{output power}}{\text{efficiency} \times \text{energy density}}$$

Assuming that the power station has an efficiency of 35% and the energy density of the coal used is 30 MJ kg^{-1}:

$$\text{mass of coal needed per second} = \frac{10^8}{0.35 \times 3 \times 10^7} = 9.5\,\text{kg}$$

$$\text{mass of coal needed per hour} = 9.5 \times 3600 \approx 3 \times 10^4\,\text{kg}$$

Formulate

Use existing knowledge to construct a precise and methodical answer to a non-mathematical problem.

Worked example

Question: Use the kinetic theory of gases to *formulate* an explanation of why the density of the Earth's atmosphere decreases with increasing distance from the Sun.

Answer: Gas molecules move at high speeds in random directions. Molecules further away from a planet's surface will have transferred kinetic energy to gravitational potential energy and slowed down. Fewer molecules have sufficient energy to reach greater heights.

Identify

This requires that you give a name for something, or select a correct answer from a number of different possibilities (which *may* be provided in the question). Only a brief answer is required.

Worked example

Question: The following equation represents the production of carbon-14 in the Earth's atmosphere by neutron bombardment of nitrogen:

$${}^{14}_{7}\mathrm{N} + {}^{1}_{0}\mathrm{n} \rightarrow {}^{14}_{6}\mathrm{C} + \mathrm{X}$$

Identify the particle denoted by X.

Answer: A proton (${}^{1}_{1}\mathrm{p}$ is an acceptable answer).

Outline

This is similar to *describe*, except that full details are *not* needed. The command signifies to the student to give only a *brief* answer, and there will usually only be 1 or 2 marks for such a question.

Worked example

Question: *Outline* how the energy carried by a water wave can be converted to electrical energy in an ocean-wave energy converter.

Answer: As the crests and troughs of the water pass through the converter, air is forced backwards and forwards past a turbine. The turbine causes coils of wire to rotate in a magnetic field and generate a voltage.

Plot

Mark the positions of points on a graph or diagram.

Construct, analyse and evaluate

This group of command terms may involve less familiar situations or a deeper understanding. They can test more complex skills, such as critical thinking and imagination.

Comment

This command term requires your judgment or opinion of a calculated numerical answer or a statement provided in a question. Usually only one comment is required.

Worked example

Question: The value for the specific latent heat of fusion of ice determined by experiment using an immersion heater was lower than the accepted value. *Comment* on this difference.

Answer: This was probably because thermal energy was transferred to the ice from the hotter surroundings, so less energy was needed from the heater to melt it.

Worked example

Question: There is an enormous amount of energy in the waves on the world's oceans. *Comment* on the fact that very little of this energy is transferred to forms that are useful to us.

Answer: The construction and maintenance costs of ocean-wave energy converters are currently much more expensive than for most other energy sources.

Deduce

This is a widely used command term. To *deduce* means to reach a conclusion (stated in the question) from the information provided. Most commonly, this will require you to show all the steps in a calculation and, in this respect, *deduce* has a similar meaning to *show*. However, *deducing* is not quite as straightforward and may involve more steps (and marks). As with *show*, it is important to show every step of the calculation.

Worked example

Question: A laser has an output power of 4.0 mW and forms a parallel beam of width 0.46 cm, which strikes a surface perpendicularly. If the wavelength of the light is 630 nm, *deduce* that photons are striking the surface at a rate of about $3.0 \times 10^{14}\,\text{cm}^{-2}\,\text{s}^{-1}$. (3 marks)

Answer:

number of photons per second in the beam $= \dfrac{\text{power}}{\text{energy of each photon}} = \dfrac{hc}{\lambda}$

$= \dfrac{4.0 \times 10^{-3}}{6.63 \times 10^{-34} \times 3.00 \times 10^{8}/6.30 \times 10^{-9}}$

$= 1.3 \times 10^{14}\,\text{s}^{-1}$

rate per $\text{cm}^2 = 1.3 \times 10^{14}/0.46 = 2.8 \times 10^{14}\,\text{cm}^{-2}\,\text{s}^{-1}$

Note that you should show the answer that is produced from the data provided, *not* just the approximate answer provided in the question.

Demonstrate

Use an example, or reasoning, to show that a proposition is correct.

Worked example

Question: Use an everyday example to demonstrate that water has a high specific heat capacity.

Answer: A glass of hot water will take a longer time to cool to room temperature, compared with similar amounts of other materials under the same circumstances.

Derive

To *derive* means to show all the physics principles and mathematical reasoning that leads to a particular equation.

Worked example

Question: *Derive* an expression for the gravitational field strength on the surface of a planet in terms of its mass M, its radius R and the universal gravitational constant G. (2 marks)

Answer: Gravitational field strength, g, is defined as $\dfrac{\text{gravitational force}}{\text{mass}}$:

$$g = \frac{F}{m}$$

From Newton's universal law of gravitation we know that:

$$F = G\frac{Mm}{R^2}$$

(In this case there is no need to explain the symbols, because they are explained in the question.)

Combining these two equation gives:

$$g = \frac{GMm/R^2}{m}$$

Determine

This command term usually relates to questions requiring numerical answers. It has a meaning very similar to *calculate*, although the term itself relates to finding a definite answer. The context of the questions may be more difficult than straightforward calculations.

Discuss

This command term requires that you present (and compare) alternative explanations and opinions, or the advantages and disadvantages of various choices.

Worked example

Question: *Discuss*, in terms of efficiency and transportability the use of natural gas rather than coal in the production of electrical energy. (3 marks)

Answer: Natural gas is preferred because (i) the conversion of the chemical potential energy in natural gas to electrical energy is more efficient than with coal, and (ii) gas can be continuously transferred along pipelines from its source to power stations, while coal has to be transferred in mobile containers of various kinds.

Answers to this kind of open-ended question can easily become too lengthy. Three marks are allocated to the answer, but the mark scheme will probably award marks to *any* three relevant comments. Note that the question requires answers *only* related to *efficiency and transportability*, so if you discuss other features (such as pollution), those comments will be ignored.

Explain

This command term is very widely used in examination papers, usually requiring you to make something understandable by giving details, or the reasons why something may, or may not, happen. The detail required in an answer can be assessed from the number of marks allocated.

Worked example

Question: A constant forward force is used to accelerate a car. *Explain* why the magnitude of the acceleration produced by a constant forward force decreases as the car moves faster. (4 marks)

Answer: Acceleration is proportional to the resultant force acting on the car. The resultant force equals the forward force less the resistive forces opposing motion. As the car moves faster, the resistive forces (mainly air resistance) increase. So the resultant force and acceleration decrease. (Four marks require that four different points of explanation are made.)

Show

This is similar to *calculate* and *determine*, but the main intention here is for you to *show in detail* how an answer (given in the question) was obtained, rather than just to perform the calculation. This kind of question may be asked in the first part of a series of calculations, so that you then have the correct data for further calculations.

Worked example

Question: An electron moved between charged parallel plates with a p.d. of 250 V across them. Show that the electric potential energy of the electron changed by 4.0×10^{-17} J.

Answer:

$$\text{potential difference} = \frac{\text{energy transferred}}{\text{charge flowing}}$$

$$V = \frac{E}{q}$$

This can be rearranged to give $E = Vq$

(Include details of *every step* of the calculation.)

So:

$$E = 250 \times 1.6 \times 10^{-19} = 4.0 \times 10^{-17}\ \text{J}$$

Sketch

This command term requires that you draw a graph, but without any numerical data. The word *sketch* does not imply 'untidy'. Your drawing *does* need to be neat, so draw using a ruler! The axes should be labelled with the quantities that they represent, as should any important features of the graph (for example gradients or intercepts). You may need to take information from another part of the question and add it to the graph.

Worked example

Question: *Sketch* a graph to show how the force between two point charges varies with their separation.

Answer: The *y*-axis should be labelled as *force*, and the *x*-axis labelled as *separation*. The origin should be labeled (0,0). An appropriate, neatly drawn and labelled curve should indicate an inverse square relationship. The graph does not need to be plotted accurately, but it should be clear that the curved line will not touch the axes.

Suggest

This command term is usually used when a single correct answer is not expected, perhaps because not enough information is available or there are many possible answers. Or a definite answer may require knowledge beyond that covered in the syllabus. Generally, there are several acceptable answers to this kind of question.

Worked example

Question: *Suggest* a reason why the melting point of ice was measured to be −1.5°C and not 0.0°C (referring to an experiment in which the temperature of some ice and water was measured over a period of time).

Answer: The thermometer used was wrongly calibrated. (Only *one* suggestion is required here. There is no way of knowing if this suggestion is actually correct. For example, an alternative answer could be that 'the ice was not pure'.)

Other terms

The following command terms are listed in the IB Diploma Physics syllabus, but they are less commonly used in examinations.

- **Analyse** – use data or information provided in a question in order to reach some kind of conclusion.
- **Compare** – describe the similarities between two or more situations or objects.
- **Construct** – describe the similarities and *differences* between two or more situations or objects.
- **Design** – this command term usually means that you are required to write some kind of plan.
- **Evaluate** – consider the advantages and disadvantages of a process.
- **Hence** – answer a part of a question by only using information from earlier in the same question.
- **Hence or otherwise** – answer a part of a question by only using information from earlier in the same question is suggested, but any other correct method will also be acceptable.
- **Justify** – give reasons to support your answer.
- **Predict** – give the expected result of a course of action or calculation.
- **Solve** – determine an answer by using algebraic methods.

Understanding mark schemes

How marks are allocated

After you have taken your IB Diploma Physics examination, your answers will be sent to the IB office in the UK. The papers are then scanned and made available on a secure IB website for examiners (based all around the world), together with an agreed *mark scheme*. Examiners must use this detailed mark scheme when carefully assessing students' work. The marking of all examiners is checked automatically by the IB's marking software to ensure that the work of all students is treated fairly and equally. Examiners know nothing about the students – except their examination number.

As already mentioned, past examination questions and their mark schemes should be an important part of your revision. You should be aware of the following points when using a mark scheme.

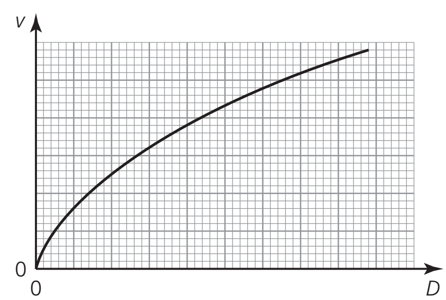
- Marking is meant to be *positive*. Answers are given credit for the understanding that they demonstrate. The examiner will *not* usually look for the exact words shown in the mark scheme, but will award marks if the same ideas are shown clearly in some other way. (Anticipated alternative answers or wording are indicated in the mark scheme by a '/'.)
- **OWTTE** means 'or words to that effect'. This is used on the mark scheme when it is expected that different students will write different acceptable answers to a certain question.
- Each marking point starts on a new line and ends with a semicolon (;).
- Occasionally for some answers, a certain word is considered to be *essential*, indicated by underlining that word in the mark scheme.
- Words in brackets (…) are used to clarify an issue for the examiners. They are not required to gain the mark.
- The separate points in a mark scheme do not need to be in any particular order.
- You will *not* be penalized for poor grammar or spelling, as long as your meaning is clear.
- Sometimes there are more relevant points (which can be made in response to a certain question) than there are marks allocated to that question. For example there may be six or more marking points for a question that only has 4 marks. In this case, *any four* of those points will result in the maximum mark of 4 being awarded.
- **ECF** means 'error carried forward'. This is used by examiners to explain why they have given marks to an incorrect answer – for example, in a calculation, the student has only got the wrong answer because they used their incorrect answer to a previous part of the same question.
- In your answers to calculations, don't forget to give a *unit* (unless it is a ratio) and use the correct number of *significant figures*. If you do not do this, you will lose a maximum of 2 marks (one for each type of mistake) for the *whole* paper.
- If you have to take a measurement from a graph on the examination paper, there will be a range of acceptable answers, but accurate measurement is required, so be careful.

Paper 3, Section A exemplar questions

Data-based questions

Note that the following past IB Physics data analysis questions are longer than those that will be included in the new examinations.

1 As part of a road-safety campaign, the braking distances of a car were measured. A driver in a particular car was instructed to travel along a straight road at a constant speed, v. A signal was given to the driver to stop and she applied the brakes to bring the car to rest in as short a distance as possible. The total distance, D, travelled by the car after the signal was given was measured for corresponding values of v. A sketch graph of the results is shown here.



a State why the sketch graph suggests that D and v are not related by an expression of the form:

$D = mv + c$

where m and c are constants. (1)

b It is suggested that D and v may be related by an expression of the form:

$D = av + bv^2$

where a and b are constants.

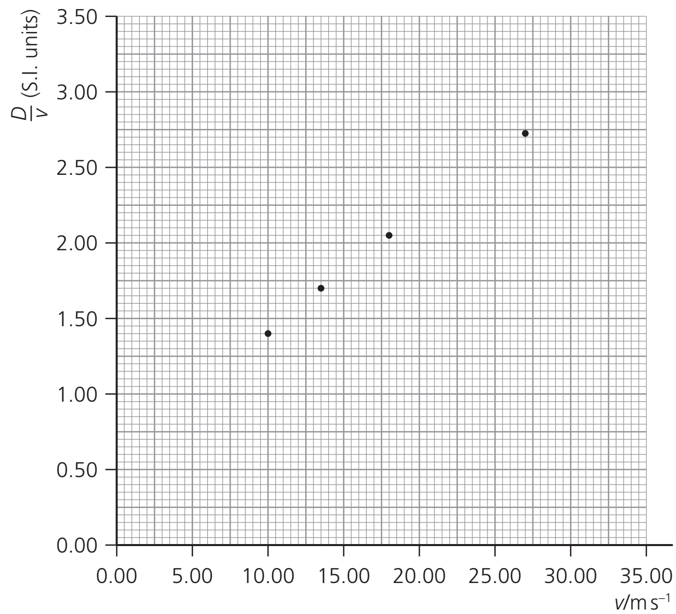
In order to test this suggestion, the data shown below are used. The uncertainties in the measurements of D and v are not shown.

$v/\text{m s}^{-1}$	D/m	$\frac{D}{v}$
10.0	14.0	1.40
13.5	22.7	1.68
18.0	36.9	2.05
22.5	52.9	
27.0	74.0	2.74
31.5	97.7	3.10

i State the unit of $\frac{D}{v}$. (1)

ii Calculate the magnitude of $\frac{D}{v}$, to an appropriate number of significant digits, for $v = 22.5\,\text{m s}^{-1}$. (1)

c Data from the table are used to plot a graph for $\frac{D}{v}$ (y-axis) against v (x-axis). Some of the data points are plotted on a graph as shown.



- **i** On a copy of the graph, plot the data points for speeds corresponding to 22.5 m s^{-1} and to 31.5 m s^{-1}. (2)
- **ii** Draw the best-fit line for all the data points. (1)

d Use your graph from (c) to determine:
- **i** the total stopping distance D for a speed of 35 m s^{-1} (2)
- **ii** the intercept on the $\frac{D}{v}$ axis (1)
- **iii** the gradient of the best-fit line. (2)

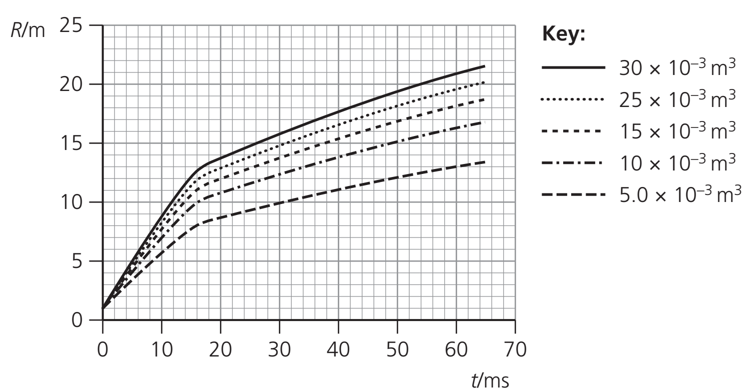
e Using your answers to (dii) and (diii), deduce the equation for D in terms of v. (1)

f The uncertainty in the measurement of the distance D is ±0.3 m and the uncertainty in the measurement of the speed v is ±0.5 m s^{-1}.
- **i** For the data point corresponding to v = 27.0 m s^{-1}, calculate the absolute uncertainty in the value of $\frac{D}{v}$. (2)
- **ii** Each of the data points in (b) was obtained by taking the average of several values of D for each value of v. Suggest what effect, if any, the taking of averages will have on the uncertainties in the data points. (2)

2 The question is about investigating a fireball caused by an explosion.

When a fire burns in a confined space, the fire can sometimes spread very rapidly in the form of a circular fireball. Knowing the speed with which these fireballs can spread is of great importance to fire-fighters. In order to be able to predict this speed, a series of controlled experiments was carried out in which a known amount of petroleum (petrol) stored in a can was ignited.

The radius, R, of the resulting fireball produced by the explosion of some petrol in a can was measured as a function of time, t. The results of the experiment for five different volumes of petroleum are shown on the graph. (Uncertainties in the data are not shown.)



a The original hypothesis was that, for a given volume of petrol, the radius, *R*, of the fireball would be directly proportional to the time, *t*, after the explosion. State two reasons why the plotted data do not support this hypothesis. (2)

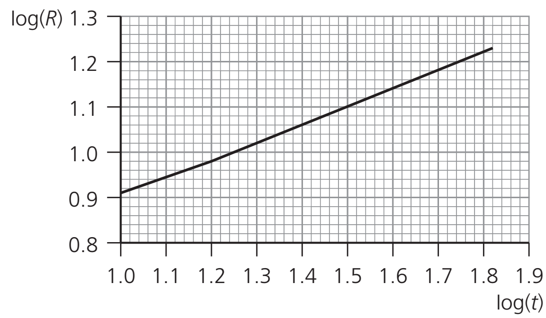
b The uncertainty in the radius is ±0.5 m. The addition of error bars to the data points would show that there is in fact a systematic error in the plotted data. Suggest one reason for this systematic error. (2)

c (HL only) Because the data do not support direct proportionality between the radius *R* of the fireball and time *t*, a relation of the form:

$$R = kt^n$$

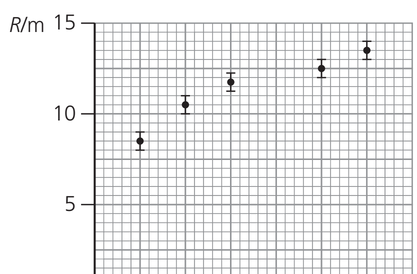
is proposed, where *k* and *n* are constants.

In order to find the value of *k* and *n*, log(*R*) is plotted against log(*t*). The resulting graph, for a particular volume of petrol, is shown. (Uncertainties in the data are not shown.)



Use this graph to deduce that the radius *R* is proportional to $t^{0.4}$. Explain your reasoning. (4)

d It is known that the energy released in the explosion is proportional to the initial volume of petrol. A hypothesis made by the experimenters is that, at a given time, the radius of the fireball is proportional to the energy, *E*, released by the explosion. In order to test this hypothesis, the radius, *R*, of the fireball 20 ms after the explosion was plotted against the initial volume, *V*, of petrol causing the fireball. The resulting graph is shown.



The uncertainties in *R* have been included. The uncertainty in the volume of petrol is negligible.

i Describe how the data for the above graph are obtained from the graph in (a). (1)

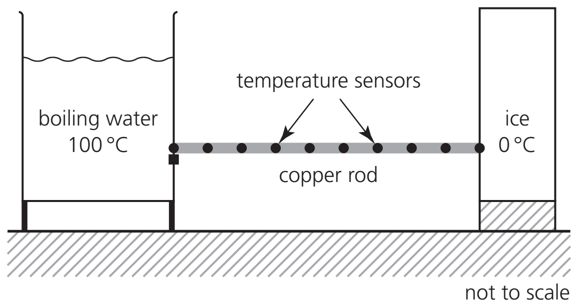
ii Make a copy of the graph and draw the line of best fit for the data points. (2)

iii Explain whether the plotted data together with the error bars support the hypothesis that R is proportional to *V*. (2)

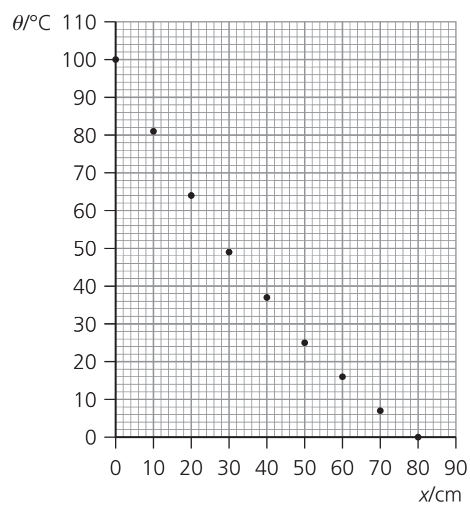
e Analysis shows that the relation between the radius *R*, energy *E* released and time *t* is in fact given by $R^5 = Et^2$.

Use data from the graph in (d) to deduce that the energy liberated by the combustion of $1.0 \times 10^{-3}\,\text{m}^3$ of petrol is about 30 MJ. (4)

3 This question is about thermal energy transfer through a rod.
A student designed an experiment to investigate the variation of temperature along a copper rod when each end is kept at a different temperature. In the experiment, one end of the rod is placed in a container of boiling water at 100°C and the other end is placed in contact with a block of ice at 0.0°C, as shown in the diagram.



Temperature sensors are placed at 10 cm intervals along the rod. The final steady state temperature, θ, of each sensor is recorded, together with the corresponding distance, x, of each sensor from the hot end of the rod. The data points are shown plotted on the axes below.



The uncertainty in the measurement of θ is ±2°C. The uncertainty in the measurement of x is negligible.

a Make a copy of the graph and draw the uncertainty in the data points for x = 10 cm, x = 40 cm and x = 70 cm. (2)

b Draw the line of best fit for the data. (1)

c Explain, by reference to the uncertainties you have indicated, the shape of the line you have drawn. (2)

d **i** Use your graph to estimate the temperature of the rod at x = 55 cm. (1)

ii Determine the magnitude of the gradient of the line (the temperature gradient) at x = 50 cm. (3)

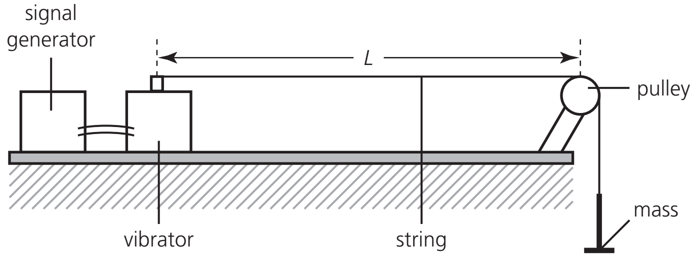
e The rate of transfer of thermal energy, R, through the cross-sectional area of the rod is proportional to the temperature gradient $\Delta\theta/\Delta x$ along the rod. At x =10 cm, R = 43 W and the magnitude of the temperature gradient is $\Delta\theta/\Delta x$ = 1.81°C cm^{-1}. At x = 50 cm the value of R is 25 W.

Use these data and your answer to (dii) to suggest whether or not the rate R of thermal energy transfer is in fact proportional to the temperature gradient. (3)

f (HL only) It is suggested that the variation of x with the temperature θ is of the form $\theta = \theta_0 e^{-kx}$ where θ_0 and k are constants.

State how the value of k may be determined from a suitable graph. (2)

4 This is a data analysis question. The frequency, f, of the fundamental vibration of a standing wave of fixed length is measured for different values of the tension, T, in the string, using the apparatus shown.

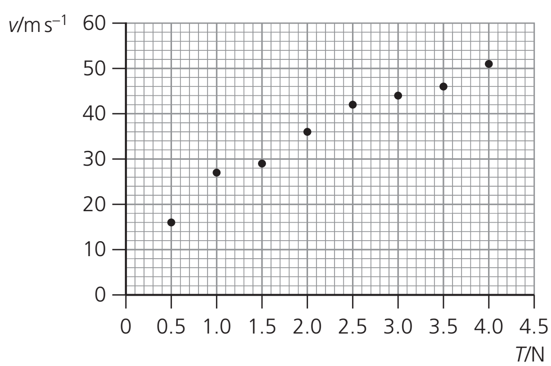


In order to find the relationship between the speed, v, of the wave and the tension, T, in the string, the speed, v, is calculated from the relationship:

$v = 2fL$

where L is the length of the string.

The data points are shown plotted on the axes below. The uncertainty in v is ±5 m s^{-1} and the uncertainty in T is negligible.



a Make a copy of the graph and draw error bars on the first and final data points to show the uncertainty in speed v. (1)

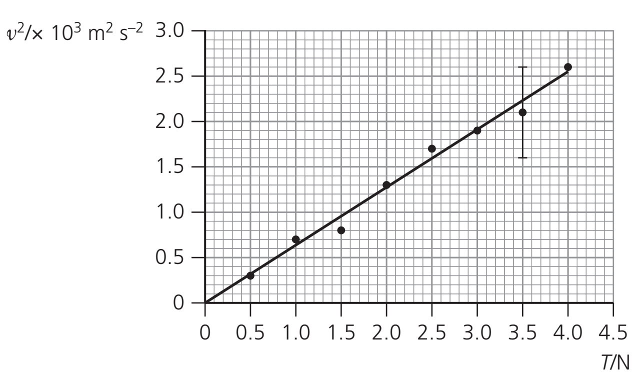
b The original hypothesis is that the speed is directly proportional to the tension T. Explain why the data do **not** support this hypothesis. (2)

c It is suggested that the relationship between speed and tension is of the form:

$v = k\sqrt{T}$

where k is a constant.

To test if the data support this relationship, a graph of v^2 against T is plotted as shown below.



The best-fit line shown takes into account the uncertainties for each data point. The uncertainty in v^2 for $T = 3.5\,\text{N}$ is shown as an error bar on the graph.

i State the value of the uncertainty in v^2 for $T = 3.5\,\text{N}$. (1)

ii At $T = 1.0\,\text{N}$ the speed $v = 27 \pm 5\,\text{m}\,\text{s}^{-1}$. Calculate the uncertainty in v^2. (3)

d Use the graph in (c) to determine k without its uncertainty. (4)

© IB Organization

Questions on experimental work

Questions about experimental work may concentrate on ten key experiments:

- Determining the acceleration of free-fall.
- Applying the calorimetric techniques of specific heat capacity or specific latent heat.
- Investigating at least one gas law.
- Investigating the speed of sound.
- Determining refractive index.
- Investigating one or more of the factors that affect resistance.
- Determining internal resistance.
- Investigating half-life experimentally or by simulation.
- Investigating Young's double-slit.
- Investigating a diode bridge rectification circuit.

1 A student uses observations of a double-slit interference pattern in a darkened room to determine the wavelength of the light used in the experiment.

a Explain why light from a laser is a good choice for this experiment.

b Discuss if it is a good idea to move the screen as far as possible from the slits.

c Explain what effect the diffraction pattern of light through the individual slits has on the observations.

2 **a** A student is investigating the resistances of wires of different cross-sectional areas of the same metal alloy (of the same length). She measured the radius of one wire to be $0.48 \pm 0.01\,\text{mm}$ and the length of the same wire to be $86.4 \pm 0.5\,\text{mm}$. The resistance of this wire was found to be $1.1 \pm 0.1\,\Omega$. She can calculate the resistivity of the wire from the equation $\rho = RA/L$. Use the data to determine the resistivity of the material of the wire and the uncertainty in the result.

b Suggest how the uncertainty in this experiment could be reduced.

3 The half-life of a radioisotope was measured in a school laboratory. Before the experiment was carried out the background count was recorded on three occasions to be $36\,\text{min}^{-1}$, $37\,\text{min}^{-1}$ and $29\,\text{min}^{-1}$.

a Explain why these readings are different from each other.

b List two safety precautions that should be taken when doing this investigation.

c The experiment had to be completed within 10 minutes. Suggest a suitable half-life for the chosen radioisotope.

Answers to self-assessment questions in Chapters 13 to 16

13 Relativity

1 805 m, 23 s
3 b 1.8×10^6 m c $500\,m\,s^{-1}$
4 a c
b $1.1c$
5 $1.2c$
6 $1.3c$
7 $8.85 \times 10^{-12}\,C^2\,N^{-1}\,m^{-2}$, $4\pi \times 10^{-7}\,T\,m\,A^{-1}$
9 a inertial
b inertial
c almost inertial
d not inertial
e almost inertial
10 Rachel, in the rocket, sees the version on the left; Gavin in the gas cloud sees the version on the right.
11 $v\Delta t$
12 $(c^2 + v^2)^{1/2}$
13 $(c^2 + v^2)^{1/2}\Delta t$
14 $c\Delta t$
15 $\Delta t = \Delta t'/(1 - v^2/c^2)^{1/2}$
17 the electron's
18 fixed point
19 yes, the observer travelling with the rod
20 no, neither observer measures proper time
21 laboratory reference frame
22 observer travelling with the rod
23 observer in laboratory reference frame
24 no, neither observer measures proper length
25 a 1.005
b 1.5
c 2.3
d 3.2
27 a 0.00
b $0.494c$
c $0.866c$
d $0.968c$
28 a 1.15
b 12 ly, 12 y
29 82 000 ly
30 410 years
31 2.3, 0.0 m, 4.2×10^{-9} s
32 2.6 m, 2.1×10^{-8} s
33 a c
b c
34 $0.94c$
35 $0.98c$
36 $0.79c$
37 $0.70c$
38 $0.9999995c$
39 c
40 $-0.17c$
41 0.600 m
42 4.2×10^{-9} s
43 a 0.96 m
b 2.7×10^{-9} s
44 a $0.65c$ b 5.9×10^{-7} s
c 114 m d $0.65c$
45 680 m
46 a 2.69×10^{-5} s
b 3.80×10^{-6} s
c The half-life of muons is only 1.5×10^{-6} s so these is a significant drop in the muon count rate between 8.00 km altitude and at the Earth's surface. If the drop in muon count rate is consistent with just over 2.5 half-lives this supports relativity; if it is consistent with 18 half-lives it is consistent with Newtonian physics.
48 a

Event	Coordinates in S (x, ct)	Coordinates in S' (x', ct')
A	(0, 0)	(0, 0)
B	(1, 1)	(0.4, 0.4)
C	(1.6, 1.1)	(1,0)
D	(2, 1.4)	(1.2, 0)
E	(1.1, 1.6)	(0, 1)
F	(2, 2.2)	(0.6, 1)
G	(2.7, 2.7)	(1, 1)

b Order of events according to an observer in S: A, B, C, D, E, F, G.

Order of events according to an observer in S': ACD (simultaneously), B, EFG (simultaneously).

49 a $0.58c$ b 1.2
c 17 years d 17 ly
e 1.2 f S'
g 17 ly/1.2 = 14 ly
50 a $0.90c$
b 2.3
c 0.44 m
d 1.2
54 a 1.60×10^{-19} J
b 6.25×10^{18} eV
c 9.00×10^{16} J
d $5.6 \times 10^{35}\,eV\,c^{-2}$
e $1.9 \times 10^{27}\,eV\,c^{-1}$
55 32.5 MW
57 a $0.86c$
b 1.9×10^{-8} s
c 9.8×10^{-9} s
58 $0.999999991c$
59 953 MeV
60 $97.8\,GeV\,c^{-2}$, 96.9 GeV, $97.8\,GeV\,c^{-1}$
61 $61\,GeV\,c^{-1}$
62 0.05175 c, $193.1\,MeV\,c^{-1}$
63 $5 \times 10^{19}\,eV\,c^{-1}$
64 110 nm
65 $100\,keV\,c^{-1}$
66 $511\,keV\,c^{-1}$
67 a 8 MeV and 542 MeV
b 1.5×10^{-13} m and 2.3×10^{-15} m
68 $4.67 \times 10^{-6}\,kg\,m\,s^{-1}$
69 $7.09\,mm\,s^{-2}$
72 a 8.32 kHz
b $7.36 \times 10^{-7}\,m\,s^{-1}$
74 a 0.3 m
b 2.1 m
c 550 m
76 Joseph Hafele and Richard Keating
77 6.2×10^{13} m
78 7.7×10^{9} m

14 Engineering physics

1 210 N
2 a 5.6 Nm
3 No. The increased torque provided by having one force further from the axis is balanced by having the other force closer to the axis.
5 $\approx 1 \times 10^{47}\,kg\,m^2$
6 a $8.1 \times 10^{-6}\,kg\,m^2$
b The ball has a constant density.
7 a If the mass of the spokes are considered negligible, it can be approximated to a thin, hollow cylinder.
b $\approx 0.1\,kg\,m^2$
8 a $0.50\,kg\,m^2$
b 2.1%
9 a $36\,kg\,m^2$
10 a $4.3 \times 10^{-3}\,rad\,s^{-1}$
b 24 minutes

11 a $1.4\,rad\,s^{-2}$
b $13\,s$
12 a $26\,rad\,s^{-1}$ b $65\,rad$
c 10 rotations d 38 rotations
13 $7.8\,rad\,s^{-2}$
14 a $314\,rad\,s^{-1}$
b $2.1\,rad\,s^{-2}$
15 $1.2\,kg\,m^2$
16 a $11\,rad\,s^{-2}$
b $9.8 \times 10^{-2}\,kg\,m^2$
17 $0.13\,rad\,s^{-2}$
18 $13.2\,rad\,s^{-1}$
21 a $-2.5\,rad\,s^{-2}$
b $35\,rad$
22 $1.4\,kg$
23 a $0.38\,rad\,s^{-1}$
24 Stars rotate about their axes and if they collapse to a very small dense core (like a neutron star), the law of conservation of angular momentum determines that their angular velocity must increase dramatically. (A full analysis would also need to take into account any loss of mass involved in the collapse.)
25 a about $8\,N\,m\,s$
b about $12\,rad\,s^{-1}$
26 $0.65\,J$
27 a $72\,rad\,s^{-1}$
b $1.9 \times 10^3\,J$
c $E_K = 2.4 \times 10^3\,J$. The two kinetic energies are similar (within 25%).
28 a $95\,cm\,s^{-1}$
b $0.12\,kg\,m^2$
31 a $18\,rad\,s^{-1}$
b The angle of the slope was small enough that the ball was able to roll.
32 72°C
33 Internal energy is the total of the potential and random kinetic energies of all of the molecules in a substance. Temperature is a measure of the average random translational kinetic energy of the molecules. Thermal energy is the non-mechanical transfer of energy from hotter to colder.
34 The most work is done in an isobaric change; the least work is done during an adiabatic change.
35 $70\,J$
36 When molecules collide with the inwardly moving surface they gain kinetic energy. If the change is done quickly there is not enough time for this energy to be dissipated out of the system.
37 a by the gas
b $3000\,J$
38 a Temperature and internal energy are constant. Pressure decreases.
b The process was isothermal because there was enough time for the energy transferred from the gas during expansion to be replaced from the surroundings.
39 For the same increase in volume, an adiabatic expansion finishes at a lower pressure because the temperature falls.
40 a $2.33 \times 10^5\,Pa$
d $3.7 \times 10^5\,J$
41 $2.45 \times 10^5\,Pa$
42 a AB
b $0.26\,mol$
c $650\,K$
d $620\,J$. This is the work done by the gas in one cycle.
43 e during process d
f $340\,J$
44
b Increasing inlet temperature or decreasing outlet temperature will have the greatest effects.
c The 'waste' internal energy could be transferred to local homes etc. to help keep them warm in cold weather.
47 Internal energy remains the same; entropy increases.
48 The molecules spread out and become more disordered, so that the entropy increases.
49 It is much easier to mix things up than to separate them because they are more disordered when mixed.
50 Only a very simple process in which no significant amount of energy is dissipated. Perhaps a few swings of a good, simple pendulum.
51 First law – energy cannot be created; second law – energy is always dissipated.
52 about $160\,J\,K^{-1}$
53 a $-70\,J\,K^{-1}$
b greater than $+70\,J\,K^{-1}$
54 $58\,J\,K^{-1}$
55 $0.763\,m$
56 If the tubes were closed, the pressures above the liquid in different tubes would not be the same.
57 $30\,m$
58 a left-hand side
b $9.2 \times 10^2\,kg\,m^{-3}$
c $1.033 \times 10^5\,Pa$
59 $8.15 \times 10^5\,Pa$
61 a $4.8\,cm$
b $925\,kg\,m^{-3}$
63 Breathing in increases the volume of the body and the weight of the water displaced. This increases the upthrust.
64 Submarines have *ballast tanks*, which can be filled with variable amounts of water and air. In this way the overall weight of the submarine can be changed.
65 rise
66 plimsoll line
67 a $260\,N$
b $1.1\,cm$
c $11\,N$
69 a 1.7×10^2
b $4.6\,m$
70 a $53\,cm\,s^{-1}$
b The gas behaves as an ideal fluid and is not compressed.
72 The aerofoil on the car is inverted (compared with that on a plane), so that it creates extra force downwards. This increases friction.
73 a $80\,kg$
b $35\,m\,s^{-1}$
c The flow is streamlined and the water at the turbine inlet is at atmospheric pressure.
75 a $5.7 \times 10^3\,Pa$
b $360\,cm\,s^{-1}$
c $1.4 \times 10^3\,Pa$
76 Because of the curved shape of the sail, the wind flows faster past the front of the sail. This creates a perpendicular force on the sail from behind (due to the pressure difference). This force has a component in the direction from which the wind is coming.
78 b $210\,m\,s^{-1}$
79 $1400\,kg\,m^{-3}$
80 a $1.3 \times 10^{-6}\,N$
b $130\,m\,s^{-1}$
c Movement is not streamlined at this speed.
82 Stokes's law can only be applied under limited conditions.
83 5000, turbulent
84 $2.4\,m\,s^{-1}$
85 25 times greater than normal air density
86 dimensions, shape, mass, stiffness
87 about 1 Hz
88 a For example: put it in water.
b No, because the physical properties of the tuning fork have not been changed.

89 a 512 Hz, 768 Hz, 1024 Hz etc.
b Increase the tension in the string.
90 The vibrations of the particles of sand tend to make them move from the positions of antinodes to nodes.
91 a 3.4 J
b 31
92 about 35
93 a $E_p = \frac{1}{2}k\Delta x^2$, $T = 2\pi\sqrt{m/k}$
b 0.14 J
c 5
d 1.6 s
94 12.6 (4π)
95 3.6 W
98 Change its mounting (the way it is attached to the car).

15 Imaging

1 a 40 cm
b The second lens should be 'fatter' in the middle.
c Both lenses were made from glass of the same refractive index.
2 a 12.5 D
b It is made from a material of lower refractive index.
3 5 cm
4 a converging lenses
b 67 cm
c +57 D
6 0.80 mm
7 a 21
b 52 cm
c 430
8 0.52 rad
9 a The image is 17 mm tall and 13 cm from the lens.
b 1.7
10 a The image is 14 cm tall and 86 cm from the lens.
b 0.71
11 20 cm from the lens
12 13 cm
13 a inverted, diminished, real
b By changing the distance between the lens and where the image is formed.
15 a 22 cm from lens
b −0.5
16 a 22 cm from lens
b −9
17 7.8 cm
18 9.2 D
19 b 4.0 cm
20 b 2.7
21 An upright, virtual image 20 cm from the lens; linear magnification = 5.0.
22 10 cm
23 a 3.1
b away from the object
c decreases
24 12 D
25 0.024 mm
26 a 5.8 cm from the lens
b 4.3
27 a 40 cm
b virtual, upright, $m = 0.8$
28 Combine with a converging lens of known focal length in order to form a real image.
29 3.0
32 The surfaces of more powerful lenses are more curved.
33 7.7 cm
34 The image is real, magnified (3.8 cm) and inverted.
36 a upright, virtual, magnified
b concave, 15 cm
c 30 cm
37 The final (inverted and virtual) image is 37 cm in front of the large mirror.
39 a 4.2 mm
b 31
40 a about 6×10^{-5} rad
b about 2×10^{-4} mm
c The microscope was used in normal adjustment.
42 a +150 D
b The powerful eyepiece lens will produce significant aberrations.
43 a 6.0×10^{-5} rad
b 7.2×10^{-5} m
c 4.8×10^{-3} rad
44 a 86 cm lens
b 41
c It would be brighter and have better resolution, but the same magnification. It may have more aberrations.
45 If a third converging lens is placed between the other two it can be used to invert the first image, producing an upright image for the eyepiece to magnify.
46 372
47 Keeping aberrations to an acceptable level becomes more and more difficult with larger telescopes. Maintaining the precise shape, alignment and mobility of the components becomes more difficult when they are much bigger and heavier.
49 Air will not contaminate the surface of the mirror; less maintenance; the shape will not change under the action of its own weight, strong supporting structure not needed; easier to steer in different directions.
51 82 m
52 4.1×10^{-4} rad
53 13 km
56 Factors include: security of data, speed of data transfer, convenience, flexibility, cost, maintaining the quality of the data transferred.
57 Induced emfs are proportional to the rate of change of magnetic flux. Higher frequencies will produce faster flux changes than low frequencies (all other factors remaining constant).
59 a 38°
b 69°
60 The signals can only have one of two significantly different values. Pulses representing 1s and 0s are still easily distinguished from each other even if they get distorted in transmission or storage.
61 a unchanged
62 a Where there is contact the radiation can cross the boundary and it will not change speed or refract.
b The claddings prevent core fibres touching each other. It has a similar (but lower) refractive index to the core, so that internal reflection occurs at the boundary between the core and the cladding.
63 76°
68 a −3.0 dB
b 90%
69 20 km
70 6.0×10^{-16} W
74 0.10 cm^{-1}
75 a 0.19 cm^{-1}
b 43%
76 a 6.4 W b 7.31×10^{14}
c 0.204 mm^{-1} d 3.39 mm
e increases
77 a 5.0 mm
b 1.2×10^{-2} $cm^2 g^{-1}$
78 A constant proportion of X-rays will be absorbed in equal distances because the probability of an X-ray having an interaction that results in absorption remains the same.
79 0.25 cm

80 $2.3 \times 10^{-3}\,mm^{-1}$

83 Because X-rays and gamma rays of the same energy are identical to each other (although they are emitted by different processes).

85 Without scattering, two parts of the film may have an intensity ratio of 4 to 1, but if the film also detects scattered X-rays the intensities may be in a ratio of, for example, (4 + 2) to (1 + 2), which equals 2 to 1.

86 a To increase the sharpness of the image.
   b To some extent the X-rays will spread out, be scattered and be absorbed in the air.

89 $1660\,m\,s^{-1}$

90 a $1.5 \times 10^{-6}\,s$
   b 18 cm
   c No, because the reflected pulse would be received after the emission of another pulse from the probe.

91 b 11%, no
   c $0.09994\,W\,cm^{-2}$
   d About 0.1%. The poor transmission is due to the low acoustic impedance of air relative to skin. The gel replaces the air between the skin and the transducer.

92 Resolution is limited by diffraction. X-rays have a *much* smaller wavelength than ultrasound and will not be significantly diffracted by the equipment used, or the parts of the body.

93 a The tissue/bone boundary has the greatest difference in acoustic impedances, so that the greatest percentage of waves are reflected there.
   b Because the waves travel at different speeds in different media.
   c 43 μs
   d $1.59 \times 10^{6}\,kg\,m^{-2}s^{-1}$

94 a In order to achieve maximum transmission of the signal, the impedances on either side of a boundary need to as close as possible in value.
   b If the acoustic impedance of the gel was the same as for skin ($1.99 \times 10^{6}\,kg\,m^{-2}\,s^{-1}$), then there would be no significant reflection at that boundary, but the transmission of the waves from the probe to the gel also needs to be considered. Typical gels have $z \approx 2.5 \times 10^{6}\,kg\,m^{-2}\,s^{-1}$.

95 a 40%
   b 15%

96 a Ultrasound waves will not penetrate the skull very well. The structure in the brain has small details that will not be resolvable with ultrasound.
   b X-rays may harm the baby; ultrasound is able to identify slight differences in soft tissues.

98 Very few ultrasonic waves will penetrate lungs containing air because of the mismatch of acoustic impedances.

16 Astrophysics

1 a Earth: $5.5 \times 10^{3}\,kg\,m^{-3}$; Jupiter: $1.4 \times 10^{3}\,kg\,m^{-3}$
  b Jupiter is a gaseous planet. Earth is mainly molten, with a solid crust and core.

2 a $3.0 \times 10^{4}\,m\,s^{-1}$
  b Mercury's speed is 1.6 times greater than the Earth's speed.

3 a 110 y
  b Almost certainly not visible (it would have to be much larger than Jupiter to be seen at that distance).

4 a Mercury, $3.3 \times 10^{23}\,kg$
  b By definition a planet must have 'cleared the neighborhood of its orbit'. There are other (less well-known) objects with similar size and mass to Pluto orbiting at about the same distance from the Sun, all of which may affect each other's motions.

5 Jupiter, $1.4 \times 10^{7}\,m$

6 Mercury, 0.39 AU; Uranus, 19 AU

7 a $9 \times 10^{23}\,km$
  b $3 \times 10^{10}\,pc$

8 a 4.2 ly
  b 3100 km

9 a 0.01 ly
  b about 4 km

10 $5.0 \times 10^{2}\,s$

11 a i About 8 months (using data from Table 18.1) and assuming that the planets are at their closest.
     ii About 300 000 years

12 10^{27}

14 Because 5 ly is considered to be a more manageable number than the equivalent 300 000 AU. The AU is more suited to planetary systems.

15 a $1/3600 = 2.8 \times 10^{-4}$°
   b $4.8 \times 10^{-6}\,rad$

16 a 1.8 pc
   b $5.6 \times 10^{16}\,m$
   c 5.9 ly

17 a 0.0125 arc-seconds
   b 0.41 arc-seconds
   c 0.375 arc-seconds

18 $3.5 \times 10^{17}\,m$

19 $3.8 \times 10^{27}\,W$

20 $1.5 \times 10^{11}\,m$

21 $L_A/L_B = 130$

22 7.6×10^{5} photons every second

23 $3.2 \times 10^{26}\,W$

24 a $8.1 \times 10^{19}\,m^2$
   b $2.5 \times 10^{9}\,m$

25 $1.8 \times 10^{4}\,K$

26 $9.5 \times 10^{-10}\,W\,m^{-2}$

27 $2.8 \times 10^{14}\,km$

28 100/1

29 The star has a radius 2.2 times greater than the Sun.

30 $5.1 \times 10^{-7}\,m$ (green)

31 a 4500 K
   b $1.6 \times 10^{22}\,m^2$
   c $3.6 \times 10^{10}\,m$

32 a $3.5 \times 10^{-7}\,m$
   b $1.0 \times 10^{28}\,W$
   c $7.2 \times 10^{-9}\,W\,m^{-2}$

33 $3.6 \times 10^{-7}\,m$

35 $7 \times 10^{9}\,m$

36 1.9:1

37 a $1.1 \times 10^{29}\,W$
   b It is a main-sequence star.
   c Sun
   d 12 000 K
   e The radius is approximately double.

38 The supergiant is about 10^4 times larger

39 a $1.7 \times 10^{30}\,W$
   b $1.3 \times 10^{-17}\,W\,m^{-2}$

40 a $2.3 \times 10^{30}\,W$
   b about 20 days

41 The luminosities of different supernovae (of the same type) are always the same.

47 a 0.0073
   b 0.10

48 $6.1 \times 10^{7}\,km\,h^{-1}$

49 The red-shift is 13 nm and the received wavelength is 423 nm.

50 5.87×10^{16} Hz

52 5300 km s^{-1}

53 43 Mpc

54 100 Mpc

55 a 3.8×10^4 km s^{-1}

b 550 Mpc

56 1 The red-shifts of the radiation received from receding galaxies indicates that their speed is proportional to their distance away.

2 The average temperature of a universe that started with a Big Bang would now be 2.76 K; radiation characteristic of this temperature is detected coming from all directions.

57 Hydrogen, because it is the most common element in the universe.

59 a decrease

b increase

60 a A source of radiation of known luminosity, which can be used to determine its distance from Earth.

b The uncertainties associated with other methods are too large when dealing with very distant galaxies.

61 The 'Big Crunch'

62 a 0.072

b 0.93

63 a 3.3×10^{34} kg

b i The gas molecules have greater average kinetic energy and therefore exert a greater pressure.

ii 1.4

64 a 3200 $L_\odot$

b 0.82 $M_\odot$

65 a 0.76 $M_\odot$

b 5.3×10^6 y

66 a 1.7×10^{19} kg

b 2.2×10^{-10}%

68 hydrogen, helium, carbon, oxygen and traces of other light elements

69 The fusion of elements only results in a more stable nucleus (and releases energy) if the product has a nucleon number of 62 or less.

74 a The Earth is not near the centre of the Milky Way, and when we look towards the centre there are more stars to be seen than when we look out of the galaxy.

b What we can see with only our eyes is an incredibly tiny part of the universe, so we are not making observations on the cosmic scale.

75 Because all space is expanding; no space is contracting.

77 The most distant galaxies have the greatest recession speeds and this simplified equation cannot be used if the recession speed approaches the speed of light.

78 The mass and volume of the universe are unknown (although we can estimate values for the *observable* universe).

79 The air is about 10^{26} times denser.

82 3×10^{54} kg

83 about 21% greater

85 a 8.1×10^5 m s^{-1}

b The star was moving directly towards or away from Earth at some point in its rotation.

c 2.6×10^{-17} kg m^{-3}

87 2.8

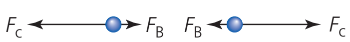
Answers to examination questions in Chapters 13 to 16

13 Relativity

Paper 3

1 a

b



c In Newtonian physics a star that is moving away from us should emit light that is travelling slower than the speed of light in a vacuum. A star that is moving towards us should emit light that is travelling faster than the speed of light in a vacuum.
The only classical alternative is that light travels through a medium called the ether but the experimental evidence suggests that this does not exist.

2 a A coordinate system/set of rulers/clocks in which measurements of distance/position and time can be made.

b i 1.25*c*
ii 1.25*c*/(1 + 0.25) shows that fraction = *c*

c Light travels at same speed for both observers. During transit time Officer Sylvester moves towards point of emission at front/away from point of emission at back. Light from front arrives first as distance is less/light from back arrives later as distance is more. Officer Sylvester observes the front lamp flashes first.
or
Time between lights arriving at Speedy is zero (according to Speedy) – (this is a proper time) so Sylvester (indeed all inertial observers) sees lights reaching Speedy simultaneously. Front lamp moving away from Speedy (according to Sylvester). Speed of light constant for all observers. Light from front lamp has to travel further to reach Speedy so must have flashed first (according to Sylvester).

d i The two events occur at the same place (in the same frame of reference)/shortest measured time.
ii 1.39×10^{-8} s

3 a i (A reference frame) in which Newton's first law holds true/that is not accelerating/that is moving with constant velocity.
ii The speed of light in a vacuum/free space is the same for all inertial observers.

b Signal from switch travels at same speed c to each lamp; but during signal transfer C_1 moves closer to/C_2 moves away from source of signal. Since speed of light is independent of speed of source, signal reaches C_1 before C_2/C_2 after C_1. According to Vladimir C_1 registers arrival of signal before C_2/C_2 registers arrival of signal after C_1.

c i 1.4 m
ii Natasha. Proper length is defined as the length of the object measured by the observer at rest with respect to the object.

d i On the return of the travelling twin according to the twin on Earth the travelling twin will have aged very little compared with himself/herself. However, since time dilation is symmetric it could be the twin on Earth who has done the least aging. The experiment suggests that it is the travelling twin who ages the least.
ii Because of the accelerations undergone by the travelling twin the situation is not symmetric/the travelling twin is not in the same inertial frame of reference/changes inertial frame of reference.

e i 890 m
ii 3200 m

f Using the laboratory half-life, most of the muons would have decayed before reaching Earth. However, many muons are detected at the surface, showing that the half-life is dilated/to the muons the distance travelled is contracted.

4 a S′ (1 mark)

b S′. This observer records the two events to occur in the same place.

c The spacetime interval between the two events is the same for both observers.
$(\Delta s)^2 = c^2(\Delta t)^2 - (\Delta x)^2 = c^2(\Delta t')^2 - (\Delta x')^2$
But $\Delta x' = 0$ while $\Delta x \neq 0$ so $\Delta t' \neq \Delta t$

d $\tan 15° \times c = 0.268c$

e 2.1 µs

f The angle of the photon worldlines is 45°. They start from Event 2 and are drawn at 90° to each other.

5 a Spacetime is the unification of the dimensions of space and time into one concept since neither is independent of the other in relativity.

b Events 0 and 1 are simultaneous in reference frame B, but are not simultaneous in reference frame A. This is because the events do not occur in the same place, and since the two reference frames are in relative motion they will observe the times that events occur differently.

c 0.42*c*

d 0.14*c*. According to A this is in the same direction as observer B/positive sign.

e The worldline drawn as a straight line from the origin and at close to 10° to the *ct*-axis.

6 a $v = 0.653c$

b $p = 809$ MeV. The protons are then accelerated in two beams travelling in opposite directions around the ring and are accelerated up to a speed of 0.97*c* before colliding.

c The worldline oscillates back and forth with constant period and is symmetrical about a vertical line. The gradient of the worldline is never less than 45°.

d $E = 7.71\,\text{GeV}$

e $u' = 0.9995c$

7 a Particle A: the total energy is the rest mass energy. Particle B: the total energy is the rest mass energy plus the kinetic energy.

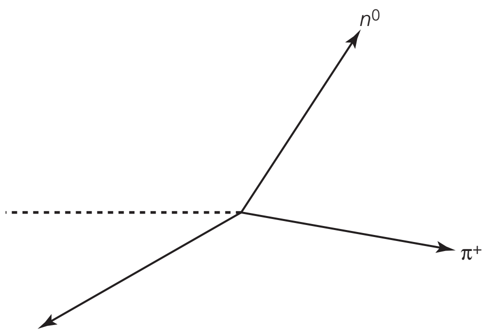
b i $0.999c$

ii $3.35\,\text{GeV}$

c i energy before collision = $6.70\,\text{GeV}$
energy of $p + n = 6.20\,\text{GeV}$

ii $p = 482\,\text{MeV}\,c^{-1}$

d



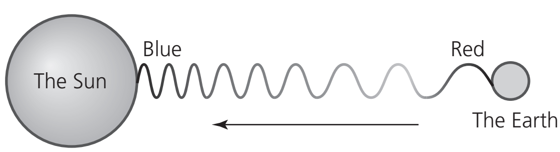
8 a According to relativity the time interval between two events will be longer for observers, in reference frames where the two events occur further apart/are moving faster relative to the events.

b i $1.47 \times 10^{-5}c$

ii It is 1.1×10^{-10} greater than 1.

c $9.50 \times 10^{-6}\,\text{s}$

d



Light appears to gain energy as it descends through a gravitational field, causing it to be blue shifted. This is because time slows down in a stronger gravitational field.

e i $1.53 \times 10^{-9}\,\text{s}$ faster per second

ii $1.32 \times 10^{-4}\,\text{s}$ faster per second

9 a i The centre is a single point to which all mass would collapse. The surface is where the escape speed is equal to c. Within this surface, mass has 'disappeared' from the universe.

ii The distance from the point of singularity to the event horizon.

iii $R_{SCH} = 3.0 \times 10^4\,\text{m}$

iv At $10^7\,\text{km}$, space is not warped, so Newtonian physics applies.

b Theory suggests that light is affected by gravitational fields *(plus diagrams or words to explain the formation of two images).*

14 Engineering physics

Paper 3

1 a An object is in rotational equilibrium if it remains stationary or continues to rotate with a constant angular velocity.

b 23 N

c 4.6 s

d The arrow must point to (or from) Q in a clockwise direction. It must have an (approximate) length such as to provide an equal and opposite torque to that provided at P. Its length must be significantly more than twice the length of *F*. The direction is not important, but the vector arrow will be shortest if it points perpendicularly to the line PQ.

2 a The moment of inertia of an object depends on the distribution of mass around the axis of rotation. A hollow sphere must have more mass at a greater distance from the axis (than a solid sphere).

b The spheres both start with the same total gravitational potential energy, which is then transferred to translational and rotational kinetic energies. The solid sphere has the smaller moment of inertia and therefore has the *lesser* rotational kinetic energy. It must have the *greater* translational kinetic energy and speed.

c 30 J

3 a Angular momentum is the product of moment of inertia and angular velocity. ($L = I\omega$)

b i The graph should be approximately sinusoidal in shape, starting at (0,0), showing two complete cycles (four maxima of angular velocity).

ii The pendulum is not an isolated system. It is acted upon by the force of gravity.

c Both $0.76\,\text{rad}\,\text{s}^{-1}$ clockwise

4 a i 0.18 mol

ii $1.9 \times 10^6\,\text{Pa}$

b i A smooth, upwardly rising curve going from A to a lower volume; followed by a vertical line to a steeper pressure; then a steeper curve back to A

ii Change b

iii The work done is equal to the area within the cycle drawn on the diagram

5 a i Adiabatic, because the change is too quick to allow the transfer of thermal energy.

ii When gas molecules collide with the inwards-moving piston they gain kinetic energy (and speed), which is measured as an increase in temperature.

b i 120 J

ii The second law states that the entropy of a system must always increase. There will be an entropy decrease in the hot source when it transfers energy, but this must be less than the sum of the two entropy increases: the gain in entropy of the sink as it absorbs energy + the gain in entropy as the engine does useful work.

6 a i $1.0 \times 10^8\,\text{Pa}$

ii The pressure due to the atmosphere is very much smaller than the pressure due to the water (approximately 1:10^3) and the question only asks for an estimate.

iii 10^4 N

b i The pressure acting upwards on the lower surfaces is greater than the pressure acting down on the upper surfaces.

ii 1.5×10^5 N

c Within the submarine there are sealed tanks that can be filled with variable amounts of water and/or air. In this way the weight of the submarine can be changed to make it greater or less than the upthrust.

7 a i A measure of a fluid's resistance to flow.

ii If the flow is streamlined (laminar, non-turbulent).

b i 3.7×10^{-5} N

ii 13 cm s^{-1}

c 53 m s^{-1}

8 b Graph should have a resonance peak at 0.7 Hz and a smaller peak at 1.4 Hz.

c Similar in shape to **b**, but smaller amplitudes.

d Resonance occurs when energy is transferred from the (infrared) radiation to the molecules of the gas if the frequency of the radiation is equal to a natural frequency at which the masses in the molecules oscillate.

9 a When the amplitudes of oscillations decrease due to resistive forces within systems.

b i under-damping (lightly damped)

ii If the amplitudes of successive peaks all have the same ratio, then the relationship is exponential. Measurements confirm that this ratio ≈ 0.5, showing that an exponential relationship is likely.

c ≈ 9

15 Imaging

Paper 3

1 a Straight rays should be drawn from the line representing the lens (from the two points where the arrows are pointing) through (and beyond) the same focal point on the principal axis. These should be labelled 'red'. Another similar pair of rays, labeled 'blue', should cross the principal axis closer to the lens.

b Red and blue light are focused in different places. This separation of colours and the resulting lack of a clear focus is called chromatic aberration.

c By using a combination of a converging and a diverging lenses made from glasses of different refractive indices.

d i 13 cm from the lens on the same side as the object.

ii 2.7

2 a See Figure 15.23, third diagram.

b i virtual, upright, diminished

ii Anywhere where a wide field of view is required, for example in a driving mirror.

3 a See Figure 15.35.

b at infinity

c See Figure 15.35.

d 2.0 cm from the lens

4 a The angle subtended at an eye by the image, divided by the angle subtended at the eye by the object.

b 3.7

c i 1.0 cm and 4.0 cm

ii 25 D lens

iii 5.0

iv 36

5 a −0.21 dB km^{-1}

b Waveguide dispersion occurs because the paths of different rays along the cable are not all exactly the same length. Material dispersion occurs because different wavelengths travel at different speeds in the glass.

c The refractive index of the glass in a graded-index fibre decreases with radial distance from the central axis. This has the effect of refracting the rays into curved paths close to the axis, so that they all travel similar distances.

6 a High-frequency pulses are sent along the coaxial cable, each as the same potential difference between the central copper wire and the surrounding copper mesh, which is earthed.

b Less attenuation; much greater data transfer rates (for cables of similar dimensions); less interference from, or to, signals in other cables; more secure.

7 a The thickness of a medium that reduces the intensity of a parallel beam of X-rays to half.

b 0.305

c A greater fraction will now be transmitted through 6.00 mm. This is because a greater half-value thickness (and smaller attenuation coefficient) means that the same thickness will absorb/scatter less.

d i X-rays are scattered in the patient's body (source too close, too broad or not collimated).

ii An oscillating collimating grid is placed between the patient and the X-ray detector/film.

e A screen placed between the patient (on the grid) and the detector contains fluorescent materials that emit light when X-rays are incident upon them.

8 a Instead of having an X-ray source and detector in fixed positions, they are rotated around the patient, who must lie very still on a bed, which is slowly moved through the beam. The whole process is computer controlled.

b i Images from CT scans have much greater resolution; the X-rays used can pass through bone.

ii Ultrasound is not believed to be any risk to health; images can be observed in real time; equipment is mobile and relatively inexpensive.

9 a between 2 MHz and 20 MHz

b In an A-scan the reflected waves received back at the probe are displayed as varying amplitudes of a p.d.–time graph. B-scans display the information in the form of varying brightness in a two dimensional real-time video image.

c Advantage: less diffraction at higher frequencies produces images with better resolution. Disadvantage: higher-frequency ultrasound waves are more attenuated than those of lower frequencies.

d 1.3

e The answer to **d** shows that there is more attenuation in the bone than in the muscle.

10 a The effect in which a system (which can oscillate) absorbs energy from another external oscillating source that is oscillating at the same frequency as the natural frequency of the system.

b When placed in a strong uniform magnetic field protons in hydrogen atoms precess around the direction of the field (at a rate known as the Larmor frequency). This frequency is in the radio frequency (RF) part of the electromagnetic spectrum. When an external RF magnetic field is applied to the patient, the protons in hydrogen atoms resonate so that their movements become in phase with each other.

c NMR does not involve sending high-energy photons into the patient (which can damage cells).

16 Astrophysics

a i nuclear fusion

ii The inwards gravitation pressure is balanced by the outwards radiation and thermal gas pressures.

b i The luminosity of a star is defined as the total power it radiates (in the form of electromagnetic waves).

ii Stars have different masses, resulting in different surface areas and temperatures.

c i The apparent brightness of a star is defined as the intensity (power/receiving area) on Earth.

ii Different luminosities; different distances from Earth.

d i $7.2 \times 10^{-8}\,\mathrm{W m^{-2}}$

ii Large stars that are relatively cool and therefore yellow/red in colour. They are not on the main sequence and they have a higher luminosity than most other stars, including red giants, because of their size. Red supergiant stars are stars that have finished their lifetime on the main sequence and will explode as supernovae to become neutron stars or black holes, depending on their mass.

iii $8.5 \times 10^{-7}\,\mathrm{m}$

2 a 2.7 K

b The average temperature of the early universe was extremely high. The Big Bang model predicts that, as it has expanded, the average temperature has fallen to the current value. (Alternatively, the wavelength originally emitted has stretched as the universe has expanded.)

c The red shift of spectral lines indicates that distant galaxies are receding at a rate which is proportional to their distance away, so that in the past they must have been closer together.

3 a Since main sequence stars will conform to the quoted equation, showing that $8 \times 10^4 = 25^{3.5}$ confirms that X is a main sequence star.

b The mass of star X is greater than Oppenheimer–Volkoff limit. This means that when it leaves the main sequence it will explode as a supernova and become a black hole.

4 a $3.8 \times 10^6\,\mathrm{m\,s^{-1}}$

b i Using the variation in luminosity of Cepheid variable stars within the cluster.

5 a Star A, because it has the lowest mass of the three, and only lower-mass stars (less than the Chandrasekhar limit) evolve into white dwarfs.

b The path must go up from A to the red giant region and then down to the white dwarf region.

6 a The cosmic scale factor, R = distance between any two points at a certain time divided by the distance between the same two points at a reference time, usually the present. So that, at the present, $R = 1$, and as the universe expands R increases. For example, at some time in the future if the size of the universe has doubled, $R \to 2$.

b 1.020

c If the expansion of the universe was linear, $R = 1.000 - 0.020 = 0.980$, but the rate of expansion was less before (because the expansion is accelerating), so a possible value could be, for example, 0.982.

7 a A sudden, unpredictable and very luminous stellar explosion.

b Type Ia supernovae occur when a white dwarf in a binary system attracts enough matter from the other star. Type II supernovae occur after the collapse of a red supergiant.

c Type Ia supernovae occur when the total mass reaches a particular, threshold value. Since this is always the same, the resulting luminosities are always the same, wherever they occur. The distance to a type Ia supernova can be determined from its known luminosity and its apparent brightness observed on Earth.

d At the very high neutron densities and temperatures in supernovae, nuclei can quickly successively capture neutrons and create heavier nuclides before there is enough time for radioactive decay.

8 a A star can only form if the temperature is high enough for fusion to occur. In order to have a high temperature the particles must have sufficient kinetic energy, which they acquire as they lose gravitational potential energy. Particles in nebulae of larger mass have greater gravitational potential energy.

b More massive stars have greater temperatures, so the rate of fusion is higher and the hydrogen is used up much quicker.

c i 0.018 × lifetime of Sun

ii It will evolve into a red giant and then a white dwarf.

9 b See Figure 16.54.

c Stars located at relatively large distances from the centre of the galaxy have rotational velocities much greater than are predicted from calculations involving the observable masses in the galaxy. There must be unobservable mass (dark matter) in the galaxy, particularly towards the edges.

d MACHOs or WIMPs (give a specific example)

10 a The universe is homogeneous and isotropic. This means that all large sections of the universe are essentially similar in structure, and observations made in any direction from any location will all be the same (on the large scale).

b Variations in observations made from Earth are due to objects relatively close to Earth (for example, the structure of our galaxy). These differences are not significant in the enormous scale of the universe.

11 a The theoretical density which would just stop the expansion of the universe after an infinite time.

b $9.2 \times 10^{-27}\,\text{kg}\,\text{m}^{-3}$

c i 5.5 nucleons m^{-3}

ii 57 cm

d i The luminosity of type Ia supernovae is known and when they occur in distant galaxies their distances from Earth can be determined. These distances are greater than would be predicted from a theory of the expansion of the universe at a constant or decreasing rate.

ii The *dark energy* theory suggests that it is present throughout the universe at very low density, providing negative pressure and thus resisting contraction.

Glossary

This glossary contains key words, equations and terms from the IB Physics Diploma course.

13 Relativity (Option A)

Standard level

This section takes words from Option A, sections A1, A2 and A3.

Dipole An object that is positively charged at one end and negatively charged at the other, or that has a magnetic north pole at one end and a magnetic south pole at the other. Magnets are always found as dipoles.

Electrical permittivity, ε The ability of different substances to transmit an electrical field. It is a measure of the amount of electrical flux (or electrical field) that is generated per unit of electrical charge in a given medium. The permittivity of free space is the permittivity of the vacuum.

Ether (or aether) A hypothetical substance, proposed as the medium through which electromagnetic waves travel. Theoretically the Earth should move through this medium and in Newtonian physics the ether should provide the only reference frame in which the speed of light in a vacuum is measured at exactly *c*, so it should define a stationary reference frame for the universe.

Event A single point in spacetime; a specific location in space at a specific moment of time. In each reference frame an event must have specific coordinates of position and time.

Galilean transformation The Newtonian, non-relativistic method of mathematically relating observations from one reference frame to another.

Inertial frame A frame of reference in which bodies that have no unbalanced forces on them obey Newton's first law, i.e. they move in straight lines with constant speed. The postulates of special relativity are valid in an inertial frame.

Inertial observer An observer who is neither accelerating nor experiencing a gravitational field.

Inertial reference frame A reference frame, or coordinate system, that is neither accelerating nor experiencing a gravitational field.

Invariant, invariant quantity The same in all reference frames. A quantity that has a value that is the same in all reference frames. In relativity examples are: the speed of light in a vacuum, spacetime interval, proper time interval, proper length, rest mass and electrical charge.

Length contraction The contraction of a measured length of an object relative to the proper length of the object due to the relative motion of an observer.

Length contraction formula $L = L_0/\gamma$, where L represents the length, L_0 represents the proper length as measured by an observer who is stationary relative to the length being measured and γ represents the Lorentz factor.

Lightline A term used by some authors to describe the worldline through spacetime made by a photon. This is normally drawn at 45° to the horizontal and vertical axes.

Lorentz factor, γ A very useful scaling factor that describes the distortion of non-invariant quantities when moving between different relativistic reference frames:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

where c is the speed of light in a vacuum and v is the relative speed of the second reference frame. The Lorentz factor ranges from close to one at classical speeds to infinity near the speed of light in a vacuum.

Lorentz transformation The mathematical formulae used to calculate the new position and time coordinates, or spatial and temporal intervals, when transferring from one relativistic reference frame to another.

Magnetic permeability, μ The ability of a substance to support a magnetic field within itself in response to an applied magnetic field. For a specific substance, permeability specifies the magnetic flux that is generated per unit current around each metre of current-carrying wire. The permeability of free space is the permeability of the vacuum.

Michelson–Morley experiment An experiment designed to measure the Earth's speed through the ether. The famous null result was the prime reason for the abandonment of the ether idea, which contributed to the development of special relativity.

Monopole A source of charge that is only positive or only negative. Magnets are never found as separate monopoles but are always found as north–south pairs called dipoles.

Muon A member of the electron family with the same charge as an electron but with 207 times the rest mass. It is therefore unstable, decaying with a half-life of 1.5×10^{-6} s, typically into an electron, a neutrino and an anti-neutrino.

Muon decay experiment A compelling experiment supporting both time dilation and length contraction. The experiment compares the levels of high-energy muons found in the atmosphere at around 10 km with those found at the Earth's surface, using the muon half-life as a means of measuring time. Classical physics predicts that the number of muons reaching the Earth's surface should be a tiny fraction of those that are formed. The measured result of around a fifth matches the predictions of relativity.

Postulates of classical or Newtonian physics A unit of time, space and mass is invariant throughout the universe. The laws of mechanics are true in all reference frames.

Postulates of special relativity The speed of light in a vacuum is the same for all inertial observers. The laws of physics are the same for all inertial observers.

Proper length The proper length of an object is the length measured by an observer who is at rest relative to the length being measured. Where the length is the distance between two events the observer must be at rest relative to a virtual object that connects the two events, so that the distance between the two events is independent of time. The proper length is always the longest length measureable by any observer; all other observers must measure a contracted length.

Proper time interval The time interval between two events as measured by an observer who records the two events occurring at the same point in space. It is the shortest time interval between events measured by any inertial observer.

Reference frame A coordinate system from which events in space and time are measured. The reference frame is commonly a set of objects that remain at rest relative to one another, from which spatial measurements can be taken, and a timing system consisting of a set of virtual clocks.

Relativistic Travelling at a significant fraction of the speed of light so that the Lorentz factor is no longer very close to 1.

Relativistic mechanics The rules that describe motion within Einstein's theory of relativity.

Rest frame The frame of reference in which a given particle or object is at rest.

Rest mass The mass of a particle (or object) at rest, or as measured by an observer who is at rest relative to the particle. According to the theory of relativity, because the energy of a particle depends on its speed, and mass and energy are equivalent, as a particle's speed increases its mass must also increase. $m = \gamma m_0$, where m_0 represents the rest mass, m represents total mass of a particle and γ is the Lorentz factor.

Scintillation An event that produces photons (light) due to the interaction of charged particles with certain materials.

Simultaneous events Events that occur at the same time in a specific reference frame, so that in this reference frame they have the same time coordinates. Events that are simultaneous in one frame may not be so in another frame.

Spacetime The combination in relativity of space and time into a single entity that is used to describe the fabric of the Universe. Fundamentally in relativity, time and space are not independent of each other and are observed differently depending on the relative motion of an observer.

Spacetime interval, Δs_2 The distance between two events across spacetime. Spacetime interval combines both the spatial and temporal elements of spacetime into a single value.

Spatial To do with the dimensions of space. A spatial interval is a length in space.

Special relativity The theory developed by Albert Einstein based on two postulates: that the laws of motion are the same for all inertial (non-accelerating) frames of reference and that the speed of light (in a vacuum) is the same for all inertial reference frames. The consequences lead to time dilation, length contraction and the equivalence of mass and energy

Synchronized Two clocks are said to be synchronized if according to an observer they are reading the same time.

Temporal To do with time. A temporal interval is an interval of time.

Thought experiment An experiment that is carried out theoretically, or in the mind, rather than actually being done normally, because it clarifies one aspect of a theory or because it is logistically impossible.

Time dilation Relative to an observer who sees the two events occurring in the same place, and so measures the proper time between the two events. All other observers measure a reduction in the time interval between two events due to the events occurring at greater spatial separations. The faster an observer is moving, relative to the observer measuring proper time, the greater the time dilation.

Time dilation formula $\Delta t = \gamma \Delta t_0$, where Δt_0 represents the proper time interval as measured by an observer who sees the first and second events occur in the same place, Δt represents that time interval between the same two events as measured by any other observer, and γ represents the Lorentz factor.

Time interval The difference between two events' time coordinates as measured from a single reference frame.

Twin paradox A paradox that appears to challenge special relativity, based on the impossibility that two twins should each find that they are older than the other. One twin remains on Earth while the other travels at high speed to a distant star and returns. Both twins claim that in their own reference frame they are stationary throughout while the other twin moves, so that the paradox appears to be symmetrical. However, the situation is not symmetrical because the travelling twin has not been in an inertial reference frame throughout and will be younger when returned to Earth than the Earth-bound twin.

Velocity addition (relativistic) If a object A has velocity u and object B has velocity v when viewed from the reference frame of object C, then the velocity u' of A with respect to B, is:

$$\frac{u+v}{1+\frac{uv}{c^2}}$$

Advanced higher level

This section takes words from Option A, sections A4 and A5.

Black hole An object of so much mass and density that spacetime becomes infinitely stretched, so that light, information and particles are unable to escape. The event horizon marks the limits of a black hole.

Equivalence principle The idea that the effects of an acceleration and of a gravitational field are completely indistinguishable. This has powerful implications and demands that physics that occurs in an accelerating reference frame must also occur in a gravitational reference frame, implying that, since spacetime is curved in an accelerating reference frame, gravitational fields and therefore mass must also distort spacetime.

Escape speed The minimum speed needed to escape to an infinitely great distance from a specific point within a gravitational field.

Event horizon An imaginary spherical surface around a black hole on which the escape speed is equal to the speed of light in a vacuum. This is used to define the dimensions of the black hole. Inside the event horizon the laws of physics become uncertain, although some predictions can be made.

General theory of relativity Einstein's generalization of special relativity to include all observers, not just observers in inertial reference frames. The main implications of the theory are to describe the distortions of spacetime in accelerating reference frames and the distortion of spacetime due to the presence of mass.

Gravitational lensing The bending of light due to the curving of spacetime around massive objects. This results in tiny shifts in the apparent positions of stars close to the Sun and in distorted or multiple images of stars as they are lensed by closer galaxies.

Gravitational mass The mass calculated by measuring the weight of an object and dividing this by the gravitational field strength.
Gravitational redshift General name for the shift in the frequency or wavelength of a photon that travels up or down in a gravitational field. The effect is a redshift if the photon travels upward, a blueshift if it travels downward.
Gravitational time dilation The slowing of time in regions of intense gravity.
Gravity (relativistic interpretation) In general relativity, gravity is explained as a consequence of the curvature of spacetime induced by the presence of a massive object.
Inertial mass The mass of an object as measured by comparing an unbalanced force on an object with the object's acceleration. Inertial mass is equal to the unbalanced force divided by the acceleration.
Kinetic energy (relativistic) The quantity $E_k = (\gamma - 1)m^0c^2$. The total energy minus the rest energy.
Momentum–energy equation The relation $E^2 = m_0{}^2c^4 + p^2c^2$ between total energy and momentum.
Relativistic momentum $p = \gamma p_0 = \gamma m_0 v$. Since mass is no longer invariant, this is just the classical momentum multiplied by the Lorentz factor.
Rest energy The minimum energy needed to create a particle, $E = m_0c^2$.
Schwarzschild radius, R_S The radius of the event horizon of a simple, uncharged, non-spinning black hole. The Schwarzschild radius defines a spherical surface where the escape velocity exactly equals the speed of light in a vacuum, which is used to specify the limits of a simple black hole.
Singularity A point at which the curvature of spacetime becomes infinite, causing the laws of physics to break down. During the formation of a black hole the surface of the star collapses inwards to a point. This point is an example of a singularity.
Solar mass A unit of mass based on the mass of the Sun. 1 solar mass = 1.99×10^{30} kg
Total energy (relativistic) The combination of potential energy and kinetic energy of an object. Normally this will be the sum of the rest energy and the kinetic energy. $E = \gamma m_0c^2$

14 Engineering physics (Option B)

Standard level

This section takes words from Option B, sections B1 and B2.

Adiabatic Occurring without thermal energy being transferred into or out of a system. $Q = 0$. An idealized situation, but approximated to by rapid changes to well-insulated systems. For an adiabatic change in an ideal monatomic gas $pV^{\frac{5}{3}}$ = constant.
Analogous Describes different systems or theories that have useful similarities.
Angular acceleration, α The rate of change of angular velocity with time, $\frac{\Delta\omega}{\Delta t}$. Unit: rad s^{-2}. $\alpha = \frac{\omega_f - \omega_i}{t}$. It is related to the linear acceleration of a point on the circumference by $\alpha = \frac{a}{r}$.
Angular momentum, L Moment of inertia multiplied by angular velocity: $L = I\omega$. Unit: kg m^2 s^{-1}
Angular velocity, ω The rate of change of angular displacement with time, $\frac{\Delta\theta}{\Delta t}$. Unit: rad s^{-1}. $\omega = \frac{2\pi}{T} = 2\pi f$.
Axis of rotation Line about which an object can rotate.
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.
Compression (of a gas) Decrease in volume. Compare with *expansion*.
Conservation of angular momentum The total resultant angular momentum of a system is constant provided that no resultant external torque is acting on it.
Couple Pair of equal-sized forces that have different lines of action, but which are parallel to each other and act in opposite directions, tending to cause rotation.
Cycle (thermodynamic) A series of thermodynamic processes that return a system to its original state (for example, the Carnot cycle).
Efficiency (thermodynamic), η Useful work done/energy input. For a Carnot cycle, maximum efficiency, $\eta_{Carnot} = 1 - \frac{T_{cold}}{T_{hot}}$.
Entropy A measure of the disorder of a thermodynamic system.
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 $\Delta S = \frac{\Delta Q}{T}$. The units of entropy are J K^{-1}.
Equations of rotational motion
$\theta = \omega_i t + \frac{1}{2}\alpha t^2$
$\omega_f{}^2 = \omega_i^2 + 2\alpha\theta$
Expansion (of a gas) Increase in volume. Compare with *compression*.
First law of thermodynamics If an amount of thermal energy, $+Q$, is transferred into a system, then the system will gain internal energy, $+\Delta U$, and/or the system will expand and do work on the surroundings, $+W$: $Q = \Delta U + W$. (This is an application of the principle of conservation of energy.)
Flywheel Dense, cylindrical (usually) mass with a high moment of inertia – added to the axes of rotating machinery to resist changes of motion and/or to store rotational kinetic energy.
Heat pump A machine designed to move thermal energy in the opposite direction from its spontaneous flow (from hotter to colder).
Inertia Property of an object that resists changes of motion (accelerations).
Internal energy of an ideal gas, U The sum of the random translational kinetic energies of all the molecules. $U = \frac{3}{2}nRT$
Irreversible process A process in which *entropy* increases; all real processes are irreversible.
Isobaric Occurring at constant pressure. $\Delta p = 0$
Isothermal Occurring at constant temperature. There is no change of internal energy: $\Delta U = 0$. pV = constant. An idealized situation, but can be approximated to by slow changes.
Isovolumetric Occurring at constant volume, so that no work is done: $W = 0$

Line of action (of a force) A straight line showing the direction in which a force is applied (through the point of application).

Moment (of a force) Sometimes used as an alternative to *torque*, especially if rotation is incomplete.

Moment of inertia, I The resistance to a change of rotational motion of an object, which depends on the distribution of mass around the chosen axis of rotation. The moment of inertia of a point mass is given by $I = mr^2$. Unit: kg m^2. The moment of inertia of any real, extended mass can be determined by the addition of the individual moments of inertia of its particles. This is represented by $I = \Sigma mr^2$.

Newton's second law for angular motion $\Gamma = I\alpha$

Order and disorder (molecular) The way in which molecules are arranged, or energy is distributed, can be described (and measured) in terms of the extent of patterns and similarities (if they exist).

Piston A solid cylinder that fits tightly inside a hollow cylinder, trapping a gas (usually). Designed to move as a result of pressure differences.

Pivot Point of rotation for a lever.

Point particle A theoretical particle that does not occupy any space, such that its properties are not dependent on its shape or size.

Principle of moments If an object is in rotational equilibrium the sum of the clockwise moments equals the sum of the anticlockwise moments.

***pV* diagram** A graphical way of representing changes to the state of a gas during a thermodynamic process.

Reservoir (thermal) Part of the surroundings of a thermodynamic system that is kept at approximately constant temperature and is used to encourage the flow of thermal energy.

Rigid body An object that keeps the same shape.

Roll Rotation of an object along a surface in which the lowest point of the object is instantaneously stationary if the surface is horizontal. Requires friction. Compare with *slipping*.

Rotation Circular motion around a point or axis.

Rotational dynamics Branch of physics and engineering that deals with rotating objects.

Rotational equilibrium Describes an object that is rotating with constant angular velocity (or is stationary). Occurs when there is no resultant torque acting.

Rotational kinetic energy, $E_{K_{rot}}$ Kinetic energy due to rotation, rather than translation. $E_{K_{rot}} = \frac{1}{2}I\omega^2$

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–Planck version: when extracting energy from a heat reservoir, it is impossible to convert it all into work.

Thermal energy, Q Non-mechanical energy transferred because of a temperature difference.

Torque, Γ Product of a force and the perpendicular distance from the axis of rotation to its line of action: $\Gamma = Fr \sin \theta$. Unit: N m

Work done when a gas changes state, W Work is done *by* a gas when it expands (W is positive). Work is done *on* a gas when it is compressed (W is negative). At constant pressure $W = p\Delta V$. If the pressure changes, the work done can be determined from the area under a *pV* diagram.

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

Advanced higher level

This section takes words from Option B, sections B3 and B4.

Aerofoil The cross-sectional shape of an aircraft wing, which is designed to produce lift using the Bernoulli effect and the force from the air striking the wing. Similar shapes are used in reverse to produce down-force on cars and lift in hydrofoils (for use in water).

Archimedes's principle When an object is wholly or partially immersed in a fluid, it experiences an upthrust (buoyancy force) equal to the weight of the fluid displaced.

Atmospheric pressure p_0 Can be considered as being due to the weight of the air above an area of 1 m^2. Acts equally in all directions.

Bernoulli effect An application of the Bernoulli equation – when the speed of a fluid flowing past a surface increases, the fluid will exert less pressure.

Bernoulli equation Equation that represents the steady flow of an ideal fluid through any enclosed system. Derived by considering the conservation of energy when the fluid changes speed and/or height: $\frac{1}{2}\rho v^2 + \rho gz + p = \text{constant}$.

Buoyancy Ability of a fluid to provide a vertical upwards force (*buoyancy force*, B) on an object placed in or on the fluid. Buoyancy force, $B = \rho_f V_f g$

Continuity equation The volume of an ideal fluid passing any point in a closed system every second must be constant: Av = constant (also called the *volume flow rate*).

Damping When resistive forces act on an oscillating system, dissipating energy and reducing amplitude. Damping may be described according to its degree: over-damping (resistive forces are so large that the amplitude reduces relatively slowly and oscillations do not occur), under-damping (many oscillations occur because resistive forces are relatively small), or *critical damping*.

Damping, critical When an oscillating system quickly returns to its equilibrium position without oscillating.

Energy of an oscillator Proportional to its amplitude squared.

Fluid Substance that can flow – usually a gas or a liquid.

Fluid dynamics The study of moving fluids.

Forced oscillations Oscillations of a system produced by an external periodic force.

Frequency, driving The frequency of an oscillating force (*periodic stimulus*) acting on a system from outside. Sometimes called *forcing frequency*.

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.

Frequency response graph Graph used to show how the amplitude of a system's oscillations responds to different driving frequencies.

Hydraulic braking system Cars and other vehicles use oil in pipes and cylinders to exert large forces on the rotating wheels in order to provide braking.

Hydraulic machinery Machines that use enclosed fluids to transfer and magnify forces.
Hydrostatic equilibrium When a fluid is either at rest, or any parts of it that are moving have a constant velocity.
Hydrostatic pressure The pressure exerted at a point in a stationary fluid because of the weight of the fluid above that point. $p = \rho_f g d$. Hydrostatic pressure acts equally in all directions. If the fluid is underneath air the overall pressure can be determined from $p = p_0 + \rho_f g d$.
Hydrostatics The study of stationary fluids.
Ideal fluid A fluid that is incompressible, non-viscous and has a steady flow if moving.
Incompressible Volume cannot be decreased.
Laminar flow Idealized model of the flow of a fluid (at relatively low speeds) in which parallel layers of fluid are visualized as moving independently of each other. Sometimes called *streamlined flow*.
Pascal's principle A pressure exerted anywhere in an enclosed static liquid will be transferred equally to all other parts of the liquid.
Periodic stimulus See *resonance*
Pitot tube Used for measuring the flow speed of a fluid, or the speed of an object through a fluid. Relies on comparing the pressure in the direct flow of the fluid to somewhere else *not* in the direct flow.
Q (quality) factor A numerical representation of the degree of damping in a system. $Q = 2\pi \times$ (energy stored in oscillator/energy dissipated per cycle), or for a resonating system oscillating regularly, $Q = 2\pi \times$ resonant frequency × (energy stored in oscillator/power loss).
Resonance The increase in amplitude that occurs when a 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.
Reynold's number, *R* Number used to predict the conditions for turbulent flow. $R = \frac{v r \rho}{\eta}$ (no unit). Different Reynold's numbers apply to different situations, but as a guide if $R < 1000$ we can expect laminar flow.
Stokes's law Viscous drag acting on a smooth, spherical object undergoing streamlined flow. $F_D = 6\pi\eta r v$
Streamlined flow See *laminar flow*
Streamlines Lines that show the paths that (massless) objects would follow if they were placed in the flow of a fluid.
Terminal speed Highest speed of an object in translational equilibrium falling vertically through a fluid. Occurs when the weight of the object is equal to the viscous drag + upthrust.
Turbulent flow Non laminar flow of a fluid, which usually occurs at higher flow rates.
Upthrust Alternative name for buoyancy force.
Venturi tube Apparatus with a narrow tube in which the fluid pressure is reduced (Bernoulli effect).
Vibration Mechanical oscillation.
Viscosity Measure of a fluid's resistance to flow. Quantified by the coefficient of viscosity, η (unit Pa s).
Viscous drag, F_D Force opposing the motion of an object through a fluid because of viscosity. See *Stokes's law*.
Volume flow rate Volume of an ideal fluid passing any point in unit time. Unit $m^3 s^{-1}$
Wind tunnel Apparatus in which the flow of air past a stationary object is observed.

15 Imaging (Option C)

Standard level

This section takes words from Option C, sections C1, C2 and C3.

Aberration, chromatic Inability of a lens to bring light of different colours (coming from the same place) to the same point focus.
Aberration, spherical Inability of a lens, or mirror, with spherical surfaces to bring light (coming from the same place) to the same point focus.
Absorption (EM waves) Transfer of wave/photon energy to other forms within a medium, so that it is not transmitted or scattered.
Aerial Metallic conductor connected to an electronic circuit, which is designed to efficiently transmit or receive electromagnetic waves (usually radio waves or microwaves).
Antenna See *aerial*
Attenuation The gradual loss of intensity of a signal as it passes through a material. Attenuation (dB) = $10\log\left(\frac{I}{I_0}\right) = 10\log\left(\frac{P}{P_0}\right)$. The intensity in an optic cable is considered to vary exponentially with distance and the attenuation is usually quoted in $dB\,km^{-1}$. See *decibel scale*
Binary number Number in which each digit can only have one of two possible values (usually 0 or 1).
Centre of curvature Where mirrors and lenses are made with surfaces that are small parts of spheres, the centre of such a sphere is called the centre of curvature of the lens or mirror surface.
Cladding (of optic fibre) Layer of glass that surrounds the central core and protects it from damage. Cladding also prevents separate cores coming in contact with each other. The refractive index of the cladding must be lower than that of the core.
Co-axial cable Cable in which a central copper wire is surrounded by an insulator and then an outer copper mesh. The mesh screens the central wire from *electromagnetic noise* (interference).
Converging mirror Also known as a *concave* mirror. Rays parallel to the principal axis are converged to a real focus that is midway between the mirror and the *centre of curvature* of the mirror surface.
Critical angle, c The smallest angle of incidence that will result in total internal reflection. $n = \frac{1}{\sin c}$
Decibel (dB) scale Logarithmic scale used for comparing widely varying powers or intensities.
Digital communication Data are transferred as a signal containing only a very large number of pulses, each of which can only have one of two possible levels (0 or 1). See *binary number*.
Dioptre, D The unit of measurement of optical power. Power in dioptres = $\frac{1}{\text{focal length in metres}}$
Dispersion The spreading (in time and length) of a pulse as it travels an increasing distance. This also results in decreased amplitude. Dispersion limits the rate at which data can be transmitted. See *waveguide dispersion* and *material dispersion*.

Diverging mirror Also known as a *convex* mirror. Rays parallel to the principal axis are diverged from a virtual focus that is midway between the mirror and the *centre of curvature* of the mirror surface.

Eyepiece Lens in an optical instrument that is closest to the eye.

Far point Furthest point from the human eye that an object can be focused clearly; usually accepted to be at infinity for normal vision.

Focal length, *f* Defined as the distance between the centre of the lens (or mirror) and the focal point.

Focal point For a converging lens (or mirror) this is defined as the point through which all rays parallel to the principal axis converge after passing through the lens (or reflecting from the mirror). For a diverging lens (or mirror) the focal point is the point from which the rays appear to diverge after passing through the lens (or reflecting off the mirror).

Focus To cause radiation (especially light) to converge to, or appear to diverge from, a point, usually with the intention of forming an image.

Graded-index fibres Optic fibres that have a non-constant refractive index. The value of the refractive index is lowest at the circumference and progressively increases towards the centre. This has the effect of confining rays to curved paths close to the centre of the fibre and therefore reducing *waveguide dispersion*. See *step-index fibres*.

Image The representation of an object that our eyes and brain 'see'.

Image properties Position, magnification, whether it is upright or inverted and whether it is real or virtual.

Imaging Formation of images.

Instrumentation Scientific equipment for observing and measuring. Developments in imaging have been mainly dependent on improved instrumentation.

Interference (electronic) See *noise*

Lens Transparent material with regularly curved surfaces that can be used to form images. *Converging* (convex) lenses usually converge rays to a real image; *diverging* (concave) lenses diverge rays away from a virtual image.

Magnification, angular, M Defined as the angle subtended at the eye by the image/angle subtended at the eye by the object: $M = \theta_i/\theta_o$.

Magnification, linear, *m* Defined as height of image/height of object: $m = \frac{h_i}{h_o} = \frac{-v}{u}$ (no unit).

Magnifying glass (simple) Single converging lens used to produce a magnified, upright, virtual image of an object placed closer to the lens than the focal point. If the image is formed at the near point the angular magnification is highest: $M_{\text{near point}} = \frac{D}{f} + 1$. Alternatively the eye can be more relaxed when the image is at infinity, then $M_{\text{infinity}} = \frac{D}{f}$.

Material dispersion Dispersion in an optic fibre that is a result of different wavelengths travelling at different speeds (and refractive indices). It can be overcome by using monochromatic radiation (e.g. radiation from an infrared LED).

Microscope, compound Two converging lenses used to produce a higher magnification of a close object than is possible with a simple magnifying glass. *Normal adjustment* means that the image is formed at the near point of the observer. Angular magnification is equal to the product of the linear magnification of the objective lens and the angular magnification of the eyepiece lens.

Microscope, electron Microscope that achieves high resolution by using electrons (which have a small wavelength) instead of light.

Near point Nearest point to the eye at which an object can be focused clearly (without straining). Usually accepted to be 25 cm from a normal eye. This distance is sometimes given the symbol *D*.

Noise (electromagnetic) Unwanted and irregular e.m.f.s induced in a conductor that is transmitting a signal. They are induced if oscillating electromagnetic waves from other sources pass through the conductor. Noise is often called *interference*, but this should not be confused with the superposition effect. If different wires within the same cable affect each other it may be described as *crosstalk*.

Object The term used to describe the place(s) from which rays/waves diverge before an optical system produces an image. Objects can either be *points* or, more realistically, *extended*.

Objective Lens or mirror in an optical instrument that receives light from the object. The quality and diameter of the objective are important factors in the quality of the final image produced by the instrument.

Opaque Describes a material through which light cannot be transmitted.

Optic fibre (communication) Fibre that transfers data using a large number of pulses of (usually) infrared radiation. Such fibres have much less attenuation than copper wires and are able to transfer much more data for similar dimensions.

Parabolic reflector Used to focus a parallel beam to a point, or to produce a parallel beam from a source placed at the focus.

Power (optical), *P* The power of a lens is the reciprocal of its focal length: $P = \frac{1}{f}$. If the focal length is measured in metres, the power is in dioptres, *D*. Refraction effects are bigger in more powerful lenses. When lenses are placed close together, their combined power is equal to the sum of their individual powers.

Principal axis Imaginary straight line passing through the centre of a lens, or curved mirror, that is perpendicular to the surfaces.

Radio-astronomy The study of the universe using radio waves.

Radio interferometry telescopes Two or more synchronized radio telescopes linked together, probably in a pattern (*array*). The combined signals form an interference pattern that provides higher resolution than from a single dish.

Ray diagram Scale drawing that shows the paths of rays from an object, through an optical system to an image. Usually the paths of three rays can be predicted and these can be used to determine the properties of the image.

Real image Image formed at a place where light rays/waves converge.

'Real is positive, virtual is negative' convention The focal lengths of diverging lenses and the distances to virtual images are given negative values, so that when using the equation $m = \frac{-v}{u}$, upright (virtual images) will always have positive magnifications, and inverted (real) images will always have negative magnifications.

Resolution The ability to see detail in an image. Measured in terms of the angle subtended by two points that can just be seen as separate. Rayleigh's criterion predicts that if the angle subtended by two points > $1.22\lambda/b$, then the two points can be resolved.

Retina The surface at the back of the eye on which images are normally formed.

Scattering (EM waves) Various processes in which the directions of waves are changed as they are passing through a medium.

Signal Information transferred in a circuit or communications system.

Step-index fibres Optic fibres that have a constant refractive index, although there is a difference (step) between the refractive index of the core and the refractive index of the cladding.

Telescopes, Earth-based Telescopes situated on the Earth's surface. Also described as *terrestrial*.

Telescope, radio Telescope that forms images using the radio waves emitted from all parts of the universe. A parabolic dish aerial focuses the waves, but the resolution may be limited by the relatively long wavelengths of radio waves. See *radio interferometry telescope*.

Telescope, refracting Two lenses used to produce an angular magnification of a distant object. The image is inverted. In normal adjustment the final image is at infinity and the angular magnification, $M = \frac{f_o}{f_e}$.

Telescope, reflecting Telescope that uses a converging mirror instead of a converging lens as the objective. In a *Newtonian mounting* a plane mirror is then used to reflect rays to the eyepiece at the side. In a *Cassegrain mounting* a diverging mirror produces extra magnification and enables the observer to look in same direction as the source of light.

Telescope, satellite-borne A telescope placed on an orbiting satellite in order to overcome the effects of the Earth's atmosphere on incident radiation.

Thin lens equation Equation linking image distance, v, to object distance, u, and focal length, f: $\frac{1}{f} = \frac{1}{v} + \frac{1}{u}$. This equation is widely used, but is only valid for rays close to the principal axis of a thin lens.

Translucent A description of a medium through which electromagnetic waves are transmitted but scattering prevents the formation of images.

Transmission (EM waves) Sending waves from one place to another without absorption and/or scattering.

Transparent A description of a medium through which electromagnetic waves are transmitted without absorption and/or scattering. Images can be formed from light that has passed through transparent materials.

Twisted pair cabling Cables that carry one or more pairs of insulated wires twisted together. The twisting reduces the effects of electronic noise (interference).

Virtual image Image formed at a place from which light rays/waves appear to diverge.

Waveguide Structure designed to transfer waves along a particular route.

Waveguide dispersion Waves passing along an optic fibre can follow slightly different paths, which can have slightly different lengths. This can result in waves that started together becoming dispersed by the time they have travelled long distances. This problem can be improved by using *graded-index fibres*.

Advanced higher level

This section takes words from Option C, section C4.

A-scan Ultrasound scan that produces an amplitude (or intensity)–time graph showing reflections from the boundaries between different media in the body. A-scans enable accurate measurements of the locations and dimensions of different parts of the body.

Acoustic impedance, Z A measure of the opposition of a medium to the flow of sound through it. Calculated from $Z = \rho c$ (unit $kg\,m^{-2}\,s^{-1}$). The amount of reflection of ultrasound waves from boundaries between media depends on how their acoustic impedances compare – the bigger the difference in impedances, the higher the percentage of incident waves that are reflected.

Attenuation of X-rays Because of absorption and scattering, the intensity of a parallel beam decreases exponentially with distance, x, travelled through a medium and it can be represented by the equation $I = I_0 e^{-\mu x}$, where μ is a constant called the *linear attenuation coefficient*. Attenuation (dB) = $10 \log\left(\frac{I_1}{I_0}\right)$.

Attenuation coefficient, linear, μ Constant that represents the amount of attenuation of X-rays per unit length in a particular medium (for radiation of a specified wavelength). Usual unit: cm^{-1}. Linked to the *half-value thickness* by the equation $\mu x_{\frac{1}{2}} = \ln 2$.

Attenuation coefficient, mass Constant that represents the amount of attenuation of X-rays per unit mass. Mass attenuation coefficient = linear attenuation coefficient/density = μ/ρ. Usual unit: $cm^2\,g^{-1}$.

B-scan Ultrasound brightness scan that produces a two-dimensional real-time video image.

Charge-coupled device (CCD) Widely used component in digital imaging. Tiny CCDs record the arrival of incident electromagnetic radiation (photons) and convert it into digital data.

Collimate Create a parallel beam.

Computed tomography (CT) Computer-controlled use of X-rays rotating around a patient to obtain good resolution images (*scans*) of multiple sections of the body. Also known as CAT scans (computed axial tomography).

Contrast Difference in intensity.

Fluorescent material Material that emits visible light after some other kind of radiation has been absorbed.

Gel (ultrasound probe) Gel of suitable acoustic impedance that is applied between the probe and the skin in order to improve the transmission of ultrasound into the body.

Gradient fields As well as the very strong primary magnetic field used in NMR, a secondary variable (gradient) field that is also applied to the patient. This ensures that different planes of the body are in slightly different magnetic fields, resulting in different *Larmor frequencies*.

Half-value thickness, $x_{\frac{1}{2}}$ Defined as the thickness of a medium that will reduce the transmitted intensity of an X-ray beam to half of its previous value.

Intensifying screen Screen containing a *fluorescent material* that is used to intensify (increase the intensity of) an image formed from X-rays.

Larmor frequency Frequency of *precession* of protons around an externally applied magnetic field. Larmor frequencies occur within the radio wave section of the electromagnetic spectrum and are proportional to magnetic field strength.

Magnetic resonance imaging (MRI) Alternative name for the use of NMR in medicine.

Nuclear magnetic resonance (NMR) Medical imaging technique in which protons in hydrogen atoms are made to spin in a very strong magnetic field. Radio frequency (RF) electromagnetic radiation is then used to make the proton spins align (a resonance effect) so that they create a detectable magnetic field. When the RF radiation is turned off the changes produced enable the location of the protons to be determined.

Piezoelectric effect Certain materials acquire a potential difference across them when they are deformed. This can be used to convert mechanical oscillations into oscillating electrical signals, and oscillating electric currents into mechanical vibrations, such as in an ultrasound transducer.

Precession When a spinning object also rotates around another axis at a lower frequency.

Probe (ultrasound) Common name for an ultrasound transducer.

Pulse repetition frequency Frequency of pulses of ultrasound, which can be adjusted to allow time for the reflected wave to be received back at the probe before the next pulse is emitted.

Quality of X-rays Describes the penetrating power of an X-ray beam (which is determined by the voltage used across the X-ray tube). More penetrating X-rays are often described as 'hard'.

Relaxation (NMR) Time during which excited protons return to their previous state. This time depends on the type of tissue involved, and the information improves the image quality.

Resonance Effect in which a system (that can oscillate) absorbs energy from another external oscillating source, such that the amplitude increases as energy is transferred.

RF coils (NMR) Coils that emit and receive the radio waves involved in NMR.

Risk analysis Scientific and technological developments can have adverse effects, which may involve health risks (direct or indirect). Every effort should be made to anticipate such risks and to evaluate the advantages and disadvantages of any research, or the subsequent use of the technology.

Scan (medical) Obtain a visual representation of the interior of the body using electromagnetic waves or ultrasound.

Sharp images Images with distinct edges and high resolution.

Tomography Obtaining images of a three-dimensional object as a series of sections or 'slices'. See *computed tomography*.

Transducer Device that converts a signal from one form of energy to another; typically to or from electricity. See *probe*.

Ultrasound Sound waves that have frequencies higher than those that can be heard by humans (≈ 20 kHz).

Ultrasound scan Medical imaging that is especially useful for identifying slight density changes between different soft tissues. Safe, mobile and inexpensive, but the resolution is disappointing.

Ultrasound scan frequency Higher frequencies (shorter wavelengths) produce better image resolution, but the attenuation is greater.

16 Astrophysics (Option D)

Standard level

This section takes words from Option D, sections D1, D2 and D3.

Accelerating universe The recession speed of distant galaxies (as determined from the redshift of Type Ia supernovae) provides evidence that the rate of expansion of the universe is increasing.

Apparent brightness, *b* Intensity (power/area) of radiation received on Earth from a star unit: W m^{-2}. Related to luminosity by: $b = \frac{L}{4\pi d^2}$

Arc-second (arcsec) $\frac{1}{3600}$ of a degree.

Astronomical unit (AU) Unit of distance used by astronomers; equal to an agreed average distance between the Sun and the Earth.

Big Bang model Currently accepted model of the universe, in which matter, space and time began at a point 13.7 billion years ago and the universe has expanded ever since. Sometimes called the *Hot* Big Bang model because early temperatures of the universe were exceptionally high.

Binary star system Two relatively close stars orbiting their common centre of mass.

Black hole After a supernova, the remaining core of a red supergiant, which is too massive to form a neutron star, will become a black hole with forces of gravity so large that light cannot escape.

Blue-shift The spectra of radiation received from stars and (the relatively few) galaxies that are moving towards Earth are shifted towards shorter wavelengths.

Cepheid variable star Type of star that is very useful in determining the distance to galaxies. The luminosity of a Cepheid variable changes in a predictable way and can be estimated from a measurement of its time period. See *period–luminosity relationship*.

Chandrasekhar limit Maximum mass of a stable white star supported against gravity by electron degeneracy pressure (= 1.4 × solar mass). More massive stars will become neutron stars or black holes.

Cluster of galaxies Group of galaxies bound together by gravitational forces. See *super clusters*. (Should not be confused with a *galactic cluster*, which is a cluster of stars within a particular galaxy.)

Comet Relatively small object of ice, dust and rock that orbits the Sun, usually with a very elliptical orbit and long period. Some have 'tails' that are visible from Earth with the unaided eye when they are close to the Sun.

Constellation An area of the night sky defined and named by the pattern of visible stars it contains. The stars may appear relatively close together, but in practice they can be a long distance apart and unconnected. Compare with *stellar cluster*.

Cosmic microwave background (CMB) radiation Spectrum of electromagnetic radiation received almost equally from all directions (see *isotropic*) and characteristic of a temperature of 2.76 K. CMB radiation is evidence in support of the Hot Big Bang model.

Cosmic scale factor, *R* Used by astronomers to represent the size of the universe by comparing the distance between any two specified places (two galaxies, for example) at two different times, one of which is usually assumed to be the present. (So that the cosmic scale factor now is 1.) The distances, and the cosmic scale factor, increase with time because the universe is expanding. *R* is closely related to red-shift: $z = \frac{R}{R_0} - 1$.

Cosmology Study of the universe (cosmos).

Dark energy Unknown form of energy the existence of which has been postulated to explain the accelerating expansion of the universe. It is believed to account for about 68% of the total mass-energy in the universe.

Electron degeneracy pressure Process occurring within white dwarf stars that keeps them stable and stops them collapsing.

Elliptical In the shape of an ellipse (oval). An ellipse has two foci on its major axis.

Expansion of the universe We know that the universe has expanded since the Big Bang and that the rate of expansion is currently increasing. But past and future expansion rates are uncertain. Possibilities are often represented on a cosmic scale factor–time graph.

Galaxy A very large number of stars (and other matter) held together in a group by the forces of gravity.

Gravitational pressure (in a star) Pressure acting inwards in a star due to gravitational forces.

Hertzsprung–Russell (HR) diagram Diagram that displays order in the apparent diversity of stars by plotting the luminosity of stars against their surface temperatures.

Hubble's law The current velocity of recession (the speed at which a galaxy appears to be moving directly away from Earth), v, of a galaxy is proportional to its distance away (from Earth), d. $v = H_0 d$, where H_0 is the Hubble constant, which can be used to estimate the age of the universe: $T \approx \frac{1}{H_0}$.

Instability strip A region of the HR diagram contain pulsating, variable stars, such as Cepheid variables.

Interstellar matter Matter between the stars – mainly gases (mostly hydrogen and helium) and dust.

Isotropic Equal in all directions

Light year, ly Unit of distance used by astronomers equal to the distance travelled by light in a vacuum in 1 year.

Luminosity, *L* Total power radiated by a star (unit: W). From Chapter 8: $L = e\sigma AT^4$. (Emissivity, e, of stars is usually assumed to be 1.)

Main sequence The band of stable stars that runs from top left to bottom right on the *Hertzsprung–Russell diagram*. Most stars are located in the main sequence.

Mass–luminosity relationship More massive *main sequence* stars have high temperatures and fast fusion rates. This means that they have shorter lifetimes. The equation that indicates the approximate relationship between mass and luminosity is $L \propto M^{3.5}$.

Milky Way The galaxy in which our solar system is located.

Moons Massive objects that orbit planets.

Nebula (plural: nebulae) Diffuse 'cloud' of *interstellar matter*; mainly gases (mostly hydrogen and helium) and dust.

Neutron stars *Main sequence* stars that have a mass greater than 8 solar masses become red supergiants at the end of their time on the main sequence. Those that have an original mass less than about 40 solar masses will become very dense neutron stars after a *supernova*. They do not collapse further because of *neutron degeneracy pressure*.

Neutron degeneracy pressure Process occurring within neutron stars that keeps them stable and stops them collapsing.

Newton's model of the universe An infinite, uniform and static universe.

Nuclear fusion Process in which light nuclei join to make a heavier nucleus, with the release of energy. Nuclear fusion is the main energy source of stars.

Occam's razor If you need to choose between two or more possible theories, select the one with the fewest assumptions.

Oppenheimer–Volkoff limit Maximum mass of a stable neutron star supported against gravity by neutron degeneracy pressure ($\approx 3 \times$ solar mass). More massive stars will become black holes.

Parallax angle (stellar), *P* Half of the angle between imaginary lines from Earth to a nearby star's position (on the background of more distant stars) drawn 6 months apart.

Parsec, pc Unit of distance used by astronomers; equal to the distance to a star that has a parallax angle of one arc-second.

Period, *T* Time taken for one complete orbit (or other regularly repeating event).

Period–luminosity relationship Graph used with Cepheid variables to determine their luminosity from knowledge of the period of the oscillations of their luminosity. This enables their distance from Earth to be determined.

Planetary nebula Material emitted from the outer layers of a red giant star at the end of its lifetime. The core becomes a white dwarf star.

Planetary system A collection of (non-stellar) masses orbiting a single star.

Plasma State of matter containing a high proportion of separated charged particles (protons, ions and electrons).

Radiation pressure Pressure in a star due to radiation emitted.

Recession speed The speed with which a galaxy (or star) is moving away from Earth.

Red giant (and red supergiant) stars Relatively cool stars, so they are yellow/red in colour; their luminosity is high because of their large size. Most stars will become red giants (or red supergiants) at the end of their time on the *main sequence*.

Red-shift Displacement of a (line) spectrum towards lower frequencies. It occurs because the distance between the source and the observer is increasing. (Similar to the *Doppler effect*.) The red-shift of radiation received from distant galaxies is evidence that the universe is expanding. Quantified by the equation $z = \Delta\lambda/\lambda_0 \approx v/c$ (no unit).

Solar system The Sun and all the objects that orbit around it.

Standard candle Term used by astronomers to describe the fact that the distance to a galaxy can be estimated from a knowledge of the luminosity of a certain kind of star within it (such as a Cepheid variable or certain type of supernova).

Star Massive sphere of plasma held together by the forces of gravity. Because of the high temperatures, thermonuclear fusion occurs and radiation is emitted.

Star map Two-dimensional representation of the relative positions of stars as seen from Earth (usually either from the northern hemisphere or from the southern hemisphere).

Stellar cluster A group of stars formed from the same nebula that are relatively close together and move as a group because they are bound together by the forces of gravity. Compare with *constellation*. *Globular clusters* are approximately spherical (like a globe) because they contain many thousands of stars. *Open clusters* contain far fewer stars, so that the overall gravitational forces are less even and the cluster has an ill-defined shape.

Stellar equilibrium *Main sequence* stars are in equilibrium under the effects of *thermal gas pressure* and *radiation pressure* acting outwards against *gravitational pressure* inwards.

Stellar parallax Method of determining the distance, d, to a nearby star from measurement of its *parallax angle*: d (parsec) = $1/p$ (arc-second), where p is the parallax angle of the star.

Stellar evolutionary path Representation on the *HR diagram* of the changes to the temperature and luminosity of a star after it leaves the main sequence.

Stellar spectra The spectra that stars emit or absorb, which are used to determine the elements present.

Sun The object around which the Earth orbits. A *main sequence* star.

Super cluster (of galaxies) Group of clusters of galaxies. May be the largest 'structures' in the universe.

Supernova Sudden, unpredictable and very luminous stellar explosion. Type Ia supernovae have a known luminosity, which makes them very useful as *standard candles*.

Thermal gas pressure (in a star) Pressure in a star due to the motion of the particles within it.

Universe All existing space, matter and energy; also called the cosmos. There may be many universes.

Universe (observable) That part of our universe that we are theoretically able to observe from Earth at this time. What we can observe is limited by the age of the universe and the speed of light.

White dwarf stars Relatively hot stars, so that they are blue/white in colour, but their luminosity is low because of their small size. They are formed at the end of the main sequence lifetimes of stars of mass < 8 solar masses. The outer layers of the star are ejected as a *planetary nebula* and the inner core, which is initially extremely hot and luminous, cools and dims to become a white dwarf star.

Wien's (displacement) law Law that connects the wavelength at which the highest intensity is emitted from a star to its surface temperature (see Chapter 8): $\lambda_{max}T = 2.9 \times 10^{-3}\,\text{m K}$.

Advanced higher level

This section takes words from Option D, sections D4 and D5.

Anisotropic Varies with direction. Compare with *isotropic*.

Closed universe ($\rho > \rho_c$) A possible future in which the universe will stop expanding and then begin to contract, and eventually end as a 'Big Crunch'.

Cognitive bias When a person's judgement is incorrectly influenced by their own experiences and opinions.

COBE, Cosmic Background Explorer Satellite launched in 1989 that investigated cosmic microwave background radiation.

Cosmic microwave background radiation, fluctuations Tiny variations in the CMB (*anisotropies*), which provide evidence about the early universe.

Cosmological principle The universe is homogeneous and isotropic (on the large scale).

Cosmic scale factor–time graphs Useful way of representing possible futures of the universe, with and without *dark energy*.

Critical density (of the universe), ρ_c The theoretical density that would *just* stop the expansion of the universe after an infinite time. The equation $\rho_c = \frac{3H^2}{8\pi G}$ can be derived from classical physics theory.

Dark matter Matter that has not been detected directly because it neither emits nor absorbs radiation. The existence of dark matter is needed to explain the higher than expected rotational velocities of stars within galaxies.

Flat universe ($\rho = \rho_c$) A possible future in which the universe will approach a limiting size at infinite time.

Homogeneous All parts are similar.

Isotropic universe What we can observe is essentially the same in every direction. This implies that there is no edge and no centre to the universe.

Jeans criterion The conditions necessary for star formation – the collapse of an interstellar cloud to form a star can only begin if its mass $M > M_J$, where M_J is known as the Jeans mass.

Lifetime of a main sequence star Depends on its mass. Greater mass means a higher rate of fusion and shorter lifetime. $T \propto \frac{1}{M^{2.5}}$

MACHOs (massive astronomical compact halo objects) A general term for any kind of massive astronomical body, which might explain the apparent presence of dark matter in the universe.

Nuclear fusion of hydrogen The main source of energy in *main sequence* stars. There are three stages in the fusion of hydrogen into helium by the *proton–proton cycle*.

Nucleosynthesis The creation of the nuclei of chemical elements by fusion or neutron capture in stars. In general, the collapse of main sequence stars of greater mass will result in higher temperatures, which means that the nuclei then have higher kinetic energies, so they can overcome the bigger electric repulsive forces involved in the fusion of heavier elements.

Open universe ($\rho < \rho_c$) A possible future in which the universe will continue to expand for ever.

Planck Space Observatory Satellite launched in 2009 that had the primary aim of investigating variations in the CMB, with resolution improved on that achieved by earlier satellites.

Proton–proton cycle See *nuclear fusion of hydrogen*

r-process (rapid neutron capture) Relatively fast nucleosynthesis of heavier elements. Occurs in supernovae that have high temperatures and neutron densities, so neutron captures are more likely that beta decays.

Red-shift, cosmological Red-shift is due to the fact that the space between the source and the observer has expanded between the time when the radiation was emitted and the time when it was received. Evidence for the Big Bang model.

Red-shift, Doppler effect Red-shift due to the fact that the observer and the source are moving apart (in unchanging space).

Rotation curve Graph showing how the rotational velocity of stars varies with distance from the centre of a galaxy. Classical physics predicts that the speeds close to the centre of the galaxy are given by $v = \sqrt{\frac{4\pi G\rho}{3}}\,r$, and at longer distances $v \propto \sqrt{\frac{1}{r}}$. but theoretical curves do not agree with observations.

s-process (slow neutron capture) Relatively slow nucleosynthesis of heavy elements (but none heavier than Bi-209) in some red giants, in which neutron captures are less likely than beta decays.

Supernova, Type Ia Occurs when two stars in a binary system join together so that *electron degeneracy pressure* is no longer sufficient to prevent the collapse of the system. Because this only occurs when the system has acquired a certain (well-known) mass, the luminosities of Type Ia supernovae are always about the same and can be used as *standard candles*.

Supernova, Type II Occurs at the end of the lifetime of a red supergiant. Results in a *neutron star* or a *black hole*.

Temperature of the universe (average) Decreases as the universe expands and the cosmological scale factor increases: $T \propto \frac{1}{R}$.

WIMPs (weakly interacting massive particles) A general term for any currently undetected particles that might explain the apparent presence of dark matter in the universe.

WMAP (Wilkinson Microwave Anisotropy Probe) Satellite launched in 2001 that had the major aim of investigating variations in the CMB.

Acknowledgements

Photo credits

Chapter 13

p.8 © CERN; **p.42** © Stocktrek Images, Inc. / Alamy; **p.46** © ESA/Hubble & NASA; **p.48** © Bettmann/CORBIS

Chapter 14

p.1 © Sinibomb Images / Alamy; **p.5** *t* © Dja65 – Shutterstock; **p.5** *b*, *r* © http://www.ricardocom; **p.7** © David Ball / Alamy; **p.9** © Duomo/Corbis; **p.14** *t* © Roy Pedersen – Fotolia.com, *b* © NASA; **p.23** © REX/Stuart Forster; **p.27** © Ljupco Smokovski – Fotolia.com; **p.28** © Shelley Gazin/CORBIS; **p.30** *t* © sciencephotos / Alamy, *b* © S McTeir for Hodder Education; **p.31** © Atlantide Phototravel/Corbis; **p.32** *b* © Alaska Stock / Alamy; **p.33** © Ashley Cooper / Alamy; **p.34** © shime – Fotolia.com; **p.36** © NASA; **p.39** © Antony Nettle / Alamy; **p.44** *t* © CHRISTIAN HARTMANN/Reuters/Corbis, *b*, *l* © alexionas – Fotolia.com, *b*, *r* © Wim de Leeuw at CWI; **p.46** © Whipple Museum of the History of Science; **p.47** © lkordela – Fotolia.com; **p.48** © Peter Tsai Photography / Alamy; **p.49** © Ron Chapple Stock / Alamy; **p.52** © REX/Sipa Press; **p.53** *t* © Pavel Svoboda – Fotolia.com, *b* © patrickw – Fotolia.com

Chapter 15

p.3 © Image via History in Pictures (historyinpictures.com); **p.19** © Viorel Sima – Fotolia.com; **p.21** © castillo Bermudez – Fotolia.com; **p.22** © macrovector – Fotolia.com; **p.26** © BSIP SA / Alamy; **p.28** © Edward H. Adelson; **p.30** *l* © H. Raab via Wikimedia Commons (http://creativecommons.org/licenses/by-sa/3.0/deed.en), *r* © Monika Wisniewska – Fotolia.com; **p.31** © G Brad Lewis / Getty Images; **p.32** © waj – Fotolia.com; **p.33** © ILYA GENKIN / Alamy; **p.34** © Phil Degginger / Alamy; **p. 36** *t* © BanksPhotos / Getty Images, *b* © ra3rn – Fotolia.com; **p.40** © PAUL WOOTTON/SCIENCE PHOTO LIBRARY; **p.43** © RTimages / Alamy; **p.44** © Juice Images / Alamy; **p.50** © RAGUET H / BSIP – Superstock; **p.51** *t* © ZEPHYR/SCIENCE PHOTO LIBRARY, *b* © FALKENSTEINFOTO / Alamy; **p.52** © MIXA / Getty Images; **p.57** *l* © BSIP/UIG Via Getty Images, *r* © ultrasoundpaedia.com, *b* © akg-images / Universal Images Group / Universal History Archive; **p.59** © National High Magnetic Field Laboratory

Chapter 16

p.2 l © Jan Tyler – Getty Images, *r* © ALEX CHERNEY, TERRASTRO.COM/SCIENCE PHOTO LIBRARY; **p.3** © pixbox77 – Fotolia.com; **p.5** *t* © NASA, b © NASA; **p.6** *t* © Stocktrek Images, Inc. / Alamy, *b*, *l* © Matthew Spinelli, *b*, *r* © JOHN R. FOSTER/SCIENCE PHOTO LIBRARY; **p.7** © NASA/Ames/SETI Institute/JPL-Caltech/SCIENCE PHOTO LIBRARY; **p.8** © Digital Vision / Getty Images; **p.9** © NASA/Space Telescope Science Institute; **p.16** *t* © Cecilia Lim – Fotolia.com, *c* © Joe McNally / Getty Images, *b* © Stocktreck / Getty Images; **p.23** © Harvard College Observatory; **p.27** © laurent dambies – Fotolia.com; **p.33** © WMAP Science Team / NASA; **p.34** © epa european pressphoto agency b.v. / Alamy; **p.42** © World History Archive / Alamy; **p.43** © David A. Hardy/www.astroart.org; **p.47** © eevl – Fotolia.com; **p.50** © ESA and the Planck Collaboration

t = top, *c* = centre, *b* = bottom, *l* = left, *r* = right

Computer hardware and software brand names mentioned in this book are protected by their respective trademarks and acknowledged.

Every effort has been made to trace all copyright owners, but if any have been inadvertently overlooked the Publishers will be pleased to make the necessary arrangements at the first opportunity.

Examination questions

Examination questions have been reproduced with kind permission from the International Baccalaureate Organization.