FOR THE
IB DIPLOMA

SECOND EDITION

Physics

OPTIONS

John Allum and
Christopher Talbot

2017 EDITION

DYNAMIC LEARNING

HODDER EDUCATION

Option A 13 Relativity

13.1 The beginnings of relativity

Revised

Essential idea: Einstein's study of electromagnetism revealed inconsistencies between the theory of Maxwell and Newton's mechanics. He recognized that both theories could not be reconciled and so, choosing to trust Maxwell's theory of electromagnetism, he was forced to change long-cherished ideas about space and time in mechanics.

Reference frames

Revised

- A reference frame is a coordinate system that is used to describe the motion of an object quantitatively, that is, to determine the position and velocity of the object.
- An example is the Cartesian coordinate system, shown in Figure 13.1, where the location of an object is specified by its coordinates relative to three mutually perpendicular axes *x*, *y*, *z*. In practice, the origin of any frame of reference can be set anywhere.
- Different frames of reference could result in different descriptions of motion.
- Looking at Figure 13.2, the small truck is moving to the right. The passenger at the back of the moving truck throws a ball upward and catches it. From his frame of reference, the ball follows a vertical path. From the perspective of the ground observer, the ball follows a parabolic path.

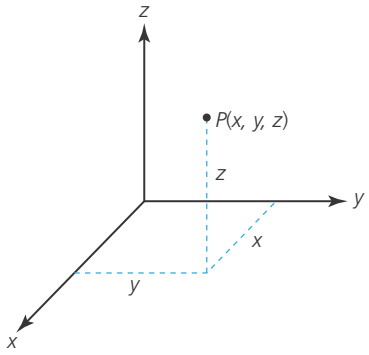


Figure 13.1

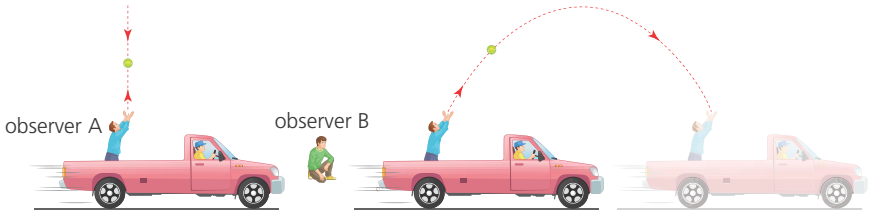


Figure 13.2 A ball thrown upward in a moving truck. (a) Observer A in the moving truck throws a ball upward and sees a straight-line path for the ball. (b) A stationary observer B sees a parabolic path for the ball.

Inertial frame of reference

- An inertial frame of reference is a frame of reference where Newton's first law of motion holds. In this frame of reference, when the net force on an object is zero, it will move in a straight line with a constant speed, as shown in Figure 13.3.
- Any reference frame moving with constant velocity with respect to an inertial frame is also an inertial frame of reference.
- An accelerating frame of reference is not an inertial frame of reference.
- A car that is cruising at a constant speed in a straight line path is an example of an inertial frame. If the car moves in a roundabout (traffic circle), it becomes a non-inertial reference frame because it is now accelerating centripetally.

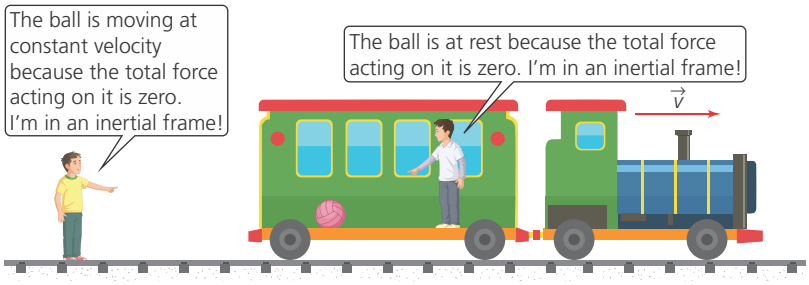


Figure 13.3 An inertial frame of reference could be at rest or moving at a constant velocity.

Newton's postulates concerning time and space

Revised

- Time is absolute.
  - Time is universal. Time flows uniformly throughout the universe. It is unaffected by any event or object. It is not possible to speed up or slow down time.
  - Once the clocks are synchronized, observers in different frames of reference will measure the same time interval for an event. Also, they would agree on the order of events and on whether or not two events happened at the same time.
- Space is absolute.
  - Space is an infinite, uniform, continuous and immovable three-dimensional matrix where one can place an object or where an event can happen.
  - Observers in different frames of reference will measure the same length or distance.
- Time and space are independent of each other. They are separate aspects of reality and are not dependent on any event or object. Events happen in space and are measured by time.
- Space and time exist without regard to any event or object. Any distance or time interval will be agreed upon by any two observers who are at rest with respect to each other or in uniform relative motion.

Galilean relativity

Revised

- Relativity is about transforming an event or a measurement from one frame of reference to another frame of reference that is moving relative to the first.
- What is now called Galilean relativity was based on the principle of invariance, that is, Newton's laws of motion hold true in all inertial frames. The velocity of the inertial frame of reference has no observable effects within the reference frame. There is no way to detect the motion of an inertial frame except by observing its motion relative to other inertial frames of reference. This makes it possible to play sports in a moving ship as shown in Figure 13.4.

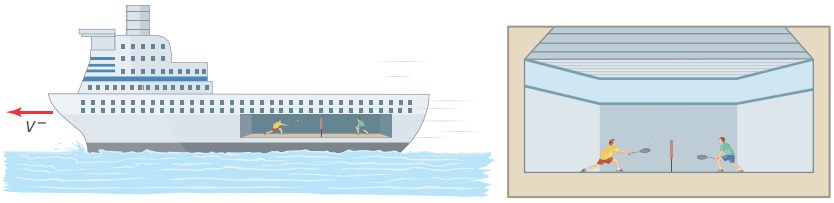


Figure 13.4 There is no difference between playing badminton inside a cruise ship moving at a constant velocity or playing badminton in a gym.

Worked example

Explain how the principle of invariance applies to this situation illustrated in Figure 13.2.

- For observer A on the truck, the ball appears to move in a vertical path with a constant downward acceleration of $9.8\,m\,s^{-2}$. Therefore, relative to his frame of reference, gravitational force acts on the ball and the equations of motion for uniform acceleration are applicable.
- To the stationary observer B, the ball is a projectile and it has a parabolic path. Its motion is affected by gravity and its position and velocity at any instant can be predicted by using the equations of motions for uniform acceleration.
- The two observers disagree on the shape of the ball's path. However, they both agree that only gravity is acting on the ball, hence the horizontal acceleration of the ball is zero and the vertical acceleration of the ball is $9.8\,m\,s^{-2}$ downwards. Also, they would both measure the same length of time that the ball was in the air. For these reasons, it can be concluded that there is no preferred frame of reference for describing the laws of mechanics.

Using the Galilean transformation equation

- The Galilean transformation equations are used to relate the position and time between two frames of reference that are moving with a uniform velocity relative to each other. In the Galilean transformation, the relative velocity between the two frames of reference is small compared to the speed of light, c, and it is assumed that length and time are absolute quantities.
- Consider two inertial frames S (x, y, z) and S′ (x', y', z') as shown in Figure 13.5. The frame S′ moves with a constant velocity, v, along the x-axis relative to S. Clocks in S and S′ were synchronized at $t = t' = 0$. For an event that happened at point P, an observer in S will describe it using coordinates (x, y, z) and time t while an observer in S′ will use coordinates (x', y', z') and time t'.

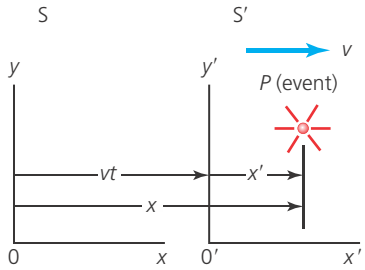


Figure 13.5

- Position transformation: from Figure 13.5, it can be seen that the coordinates of event P in S and S′ are related by the equations below.

$$x = x' + vt \qquad x' = x - vt$$
$$y = y' \qquad y' = y$$
$$z = z' \qquad z' = z$$
$$t = t' \qquad t' = t$$

- Velocity transformation. In Figure 13.6, a particle is moving with a constant velocity u' in the direction shown relative to frame S′, which is moving with

a constant velocity v relative to frame S. It is easy to see that the velocity u of particle R relative to S can be obtained by adding the two velocities v and u', that is:

$u = v + u'$

- Note that the addition is a vector addition.

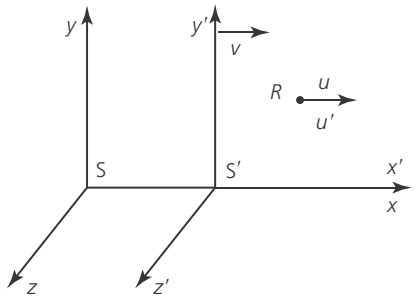


Figure 13.6 Frame S′ is moving relative to frame S with a velocity v. Particle R is moving with speed u relative to S and u' relative to frame S′.

QUESTIONS TO CHECK UNDERSTANDING

1 Concept of length in Galilean relativity. Consider the rod in Figure 13.7. In frame S, the length of the rod is $x_2 - x_1$. In frame S′ which is moving with a velocity v relative to S, the length of the rod is $x'_2 - x'_1$. Compare the length of the rod in the two frames of reference, S and S′.

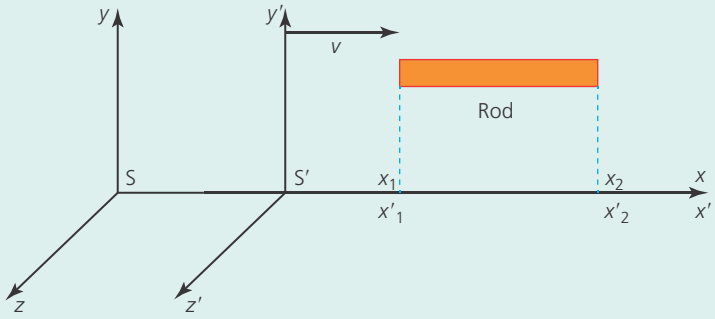


Figure 13.7

2 A ball is thrown at 10.0 m s^{-1} inside a train that is moving relative to the ground at 45 m s^{-1} (Figure 13.8). Calculate the speed of the ball relative to the ground if the ball is thrown:

a forward

b backward.

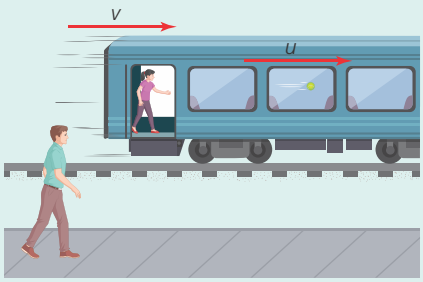


Figure 13.8

3 A 3000 kg lorry moving at 30.0 m s^{-1} collides and attaches on to a 1500 kg car initially at rest. What is their combined velocity? If the car is moving to the left, show that the momentum is conserved in an inertial frame moving at 10.0 m s^{-1} in the direction of the moving car.

NATURE OF SCIENCE

Ideas about space and time

The theory of special relativity led to a radical change in our understanding of the nature of the universe, the fundamental structure of which consists of space and time. Einstein's realization that the speed of light is a universal constant revealed that space and time are not independent of each other but are woven together in a four-dimensional continuum, referred to as 'space-time'.

Maxwell and the constancy of the speed of light

Revised

- Maxwell's greatest achievement was uniting electricity and magnetism into a single theory. The main concepts of his electromagnetic theory are summed up in his four equations which can be interpreted as follows:
  - The magnitude of an electric field is proportional to the charge present. Electric charges are monopoles. They are either positive or negative.
  - There is no experimental proof for monopoles. Magnetic poles are always found as north–south pairs.
  - An electric field can be generated by a changing magnetic field.
  - A magnetic field can be generated by an electric current or a changing electric field.
- Maxwell's theory predicted that light is an electromagnetic wave and that the speed of light in a vacuum is equal to $2.998 \times 10^8\,m\,s^{-1}$ (quoted as $3.00 \times 10^8\,m\,s^{-1}$ in the Data Booklet).
- More importantly, Maxwell's theory also proposed that the speed of light does not change in any reference frame, that is, it is invariant. The speed of light is the same whether the observer is stationary, moving towards or away relative to the light source.
- The constancy of the speed of light did not agree with Galilean relativity, where the speed of light is affected by the relative velocity between the source and observer.
- Maxwell's equations were not invariant under Galilean transformation. Since the equations were dependent on the speed of light, the transformation would give different outcomes in different reference frames, which is a violation of the principle of invariance.

NATURE OF SCIENCE

Need for consistency of theories

Maxwell's electromagnetic theory successfully unified the fields of electricity and magnetism. However, unlike Newtonian mechanics, Maxwell's theory gave inconsistent results under Galilean relativity. Einstein's desire to make the existing theories of physics consistent with each other led to the development of his theories of relativity.

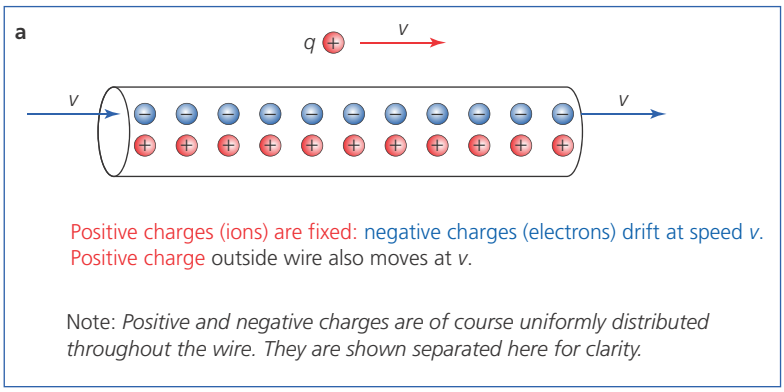
Force on a charge or current is relative to the inertial frame

Revised

Observers in different frames will agree on how an electromagnetic system behaves but will give different explanations for its behaviour. Whether the force on a charge or current is electric or magnetic in nature is relative to the frame of reference.

Determining the fields observed by different observers

- Force on a charge moving perpendicular to a magnetic field.
- Consider a positive charge, q, moving with a velocity, v, parallel to a long straight wire carrying a current.



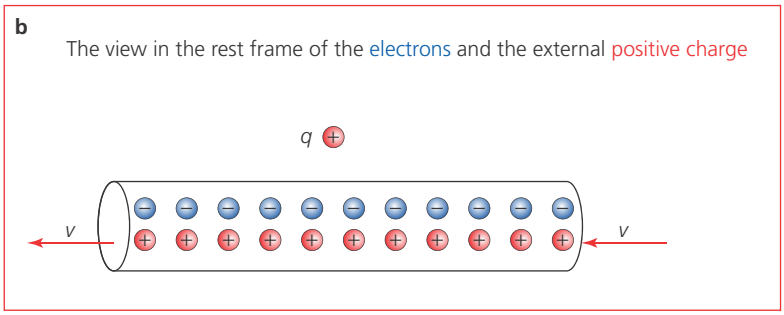


Figure 13.9 (a) In the laboratory frame, the charge q and the electrons are seen to be moving to the right. (b) In the electrons' reference frame, the charge q is at rest and the positive charges are moving to the left.

- As seen from the laboratory reference frame, there is an upward magnetic force on the charge q. The situation is shown in Figure 13.9a.
  - There is no net charge in the wire, thus there cannot be an electrostatic force on the charge q.
  - The positive charge q and the electrons are both moving to the right. The electric current in the wire generates a magnetic field that exerts an upward force on the charge q.
- As seen from the reference frame of charge q, there is an upward electrostatic force on the charge q. The situation is shown in Figure 13.9b.
  - The charge q is at rest. It cannot be affected by the magnetic force produced by the moving positive charges.
  - Inside the wire, the electrons are at rest, while the positive charges are moving to the left.
  - The relative motion of the positive charges causes a relativistic length contraction in the spacing between charges (this will be learned later). This means that there are more positive charges per unit length than there are negative charges per unit length.
  - In this case, the unequal concentration of charges produces a net positive electrical field that exerts an electrostatic force on the charge which is equal in magnitude and direction to the original magnetic force on the test charge.
- Observers in both frames of reference agree that there is a net upward force on the charge q. However, each has a different description of the nature of the force.

Force on two charged particles moving with parallel velocities:

- Consider two equal positive charges, each of charge q, moving parallel to each other with a velocity v.
- Figure 13.10a shows the situation from the reference frame of the two charged particles. The two charges are at rest relative to each other. Therefore, each particle exerts a repulsive electrostatic force on the other.

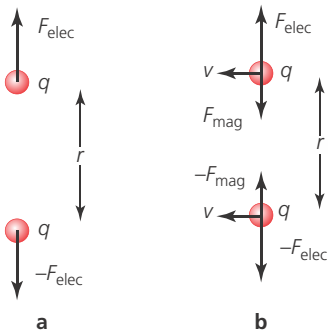


Figure 13.10

- Figure 13.10b shows the situation from a stationary reference frame.
  - In this frame, the observer sees the two electrons move with a velocity v. The motion of each electron generates a magnetic field that leads to a magnetic attraction between the two charges.
  - The motion of the charged particles will cause an increase in the electric field due to relativistic length contraction (this will be learned later). The electrostatic force thus increases.
  - Therefore, as observed, in this frame, the net force on each charge is equal to the difference between the increased electrostatic repulsion and magnetic attraction. The net force is still the same and found to be repulsive.
- Again, it can be seen from this example that different frames of reference detect different natures of force, but agree on the behaviour of the electric charge.

13.2 Lorentz transformations

Revised

Essential idea: Observers in relative 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. The 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.

The two postulates of special relativity

Revised

- **Postulate 1.** The laws of physics are the same for all inertial frames of reference.
  - The postulate implies that all laws of physics, including those of mechanics, thermodynamics, electromagnetism, quantum mechanics, etc., are the same in all inertial frames of reference. Any experiment conducted in a laboratory that is at rest will give the same results when performed in a laboratory that is moving at a constant velocity relative to the first.
  - This also means that there is no absolute rest frame of reference with which all observers in the universe would agree to be at rest at all times. Only relative positions and velocities between frames of reference are meaningful.
  - All inertial frames of reference are equally valid.
- **Postulate 2.** The speed of light in a vacuum, c, is the same in all inertial frames of reference and is independent of the motion of the light source or of the observer.
  - The speed of light c is a universal and fundamental constant. This is consistent with Maxwell's equations.
  - Note that the second postulate is consistent with the first postulate. If the speed of light were different in different inertial frames then this would make it possible to distinguish between inertial frames; and thus a preferred absolute frame could be identified, in contradiction of the first postulate.

NATURE OF SCIENCE

Thought experiments

There is no one way to do science and therefore, there is no universal step-by-step scientific method. Einstein formulated his theories of relativity by using thought experiments rather than actual experiments. A thought experiment is one that is carried out only through the imagination and the results can be reasoned out theoretically. It is usually impossible to perform, but a powerful method of describing the implications of the theory.

Clock synchronization

Revised

- Clocks are synchronized when they show the same time at the same instant. One way to synchronize clocks is to send light signals to clocks located at different points in space and start the clocks when the signal is received.
- Figure 13.11a shows how two clocks in the same inertial frame can be synchronized. A spherical flash of light is emitted from a source located exactly halfway between two clocks that are preset to zero and are not running. When the light is received by the two clocks, they start simultaneously. The two clocks are now synchronized.

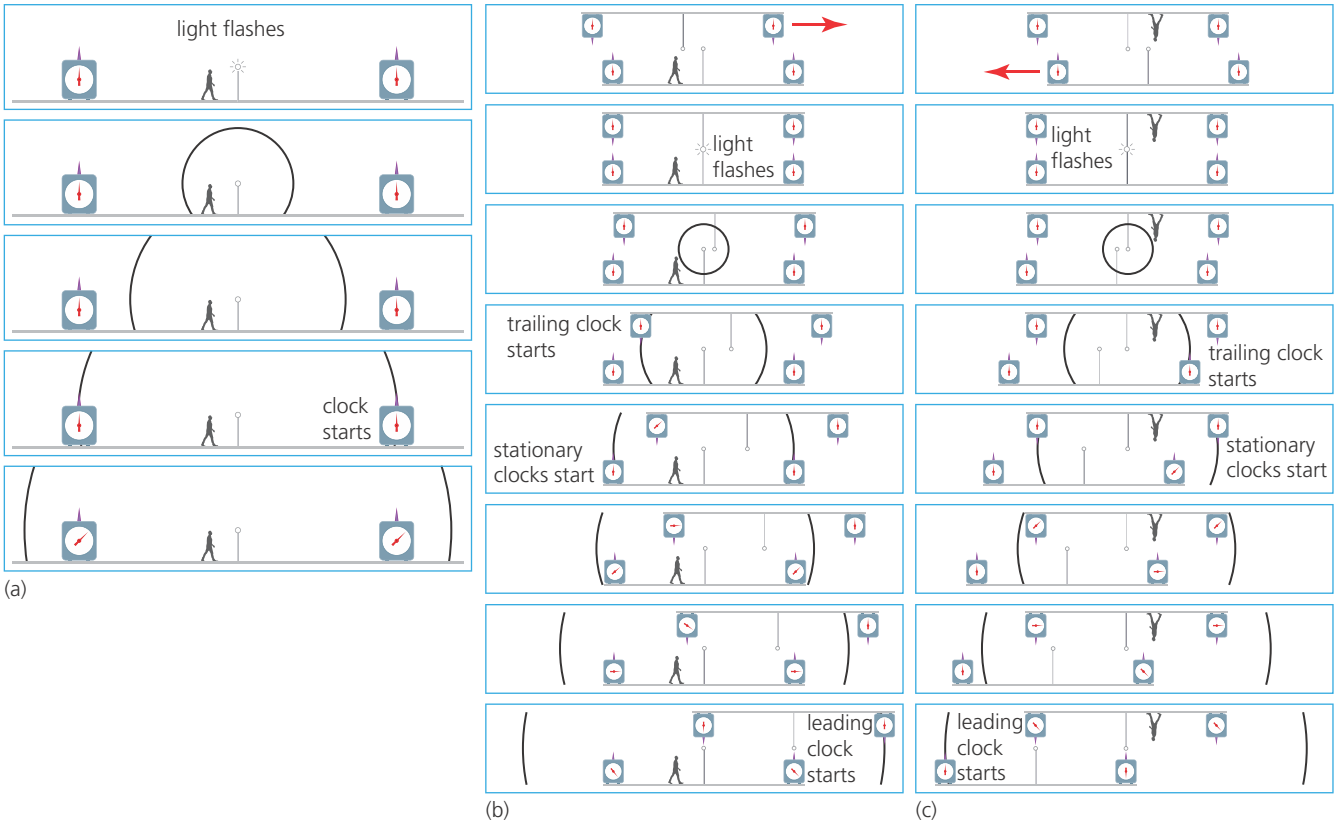


Figure 13.11 (a) Clock synchronization in a single inertial frame; (b) clock synchronization as seen by lower observer; (c) clock synchronization as seen by upper observer.

- Figure 13.11b shows an attempt to synchronize clocks in two different inertial frames of reference. In this set-up, the lower frame is at rest and the upper frame of reference is moving to the right relative to the lower frame.
  - A light pulse is emitted by a source that is stationary relative to the lower frame when the clocks are aligned to each other. As seen by the observer in the lower frame, the light pulse will reach the two clocks in his frame at the same time, thus the clocks will start simultaneously. The clocks in the lower frame are synchronized.
  - However, in the upper frame, the left clock is moving towards the pulse, while the right clock is moving away from the pulse. The light pulse reaches the left clock first and starts before the right clock does. Thus, these two clocks do not start simultaneously and are not synchronized.
- Figure 13.11c shows the same attempt to synchronize clocks, but this time from the perspective of the observer located in the upper frame of reference. He is at rest and sees the lower frame to be moving to the left. A light pulse is emitted by a source that is at rest relative to the upper frame. It is easy to see that the clocks in the upper frame are synchronized. However, in the

lower frame, the light pulse will reach the right clock first. Therefore, the clocks in the upper frame are synchronized but the clocks in the lower frame are not.

- Figure 13.11b, c demonstrates that it is not possible for an observer to synchronize clocks that are located in different inertial frames of reference.

Simultaneity is relative

Revised

The *principle of simultaneity* is explained in the following example.

- Consider two observers, Riel who is at rest relative to the ground and Rion who is in the middle of a fast moving platform. A light pulse was emitted from each end of the platform. The light pulses reached Rion simultaneously at the moment he passed by Riel, as shown in Figure 13.12.
- The two light pulses were received by Rion at the same time, but were these two light pulses emitted from each end simultaneously? Figure 13.13 shows the observations in each frame of reference.

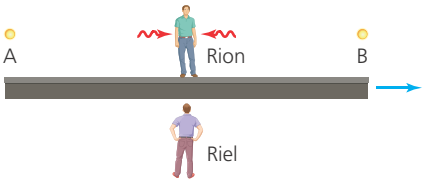
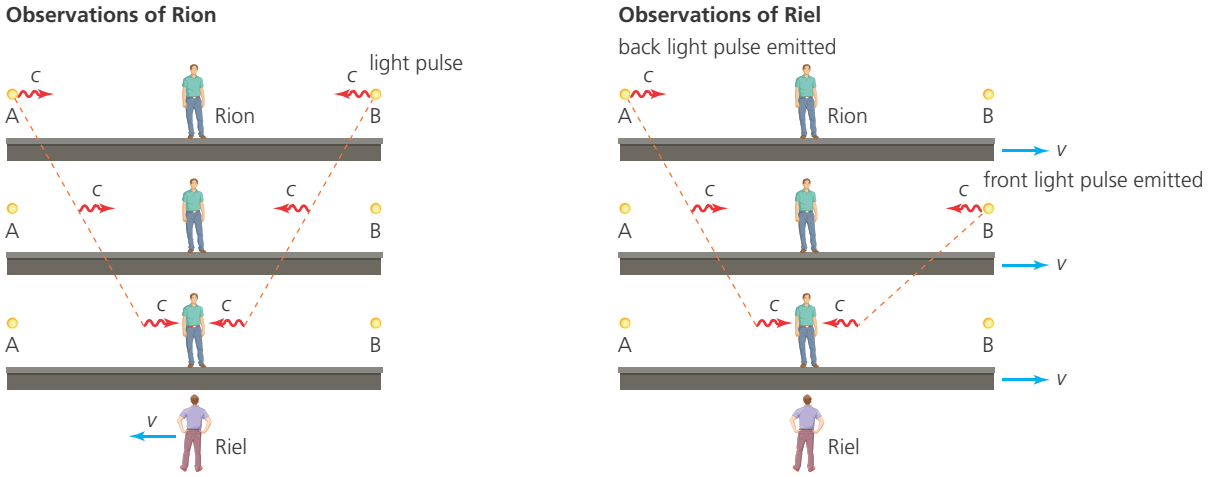


Figure 13.12



Rion's observations (on the moving platform):
- Rion received the light signals at the same time.
- Light travels at speed *c* from each end.
- Distance travelled by light from each end is the same.
- Therefore, the two light pulses were emitted simultaneously.

Riel's observations (on the ground):
- Rion received the light signals at the same time.
- Light travelled at speed *c* from each end.
- Rion moved away from the back light pulse but moved towards the front light pulse. To reach him, the light pulse from the back travelled a longer distance than that from the front.
- Therefore, the light pulse from the back end was emitted first.

Figure 13.13

- The emission of light pulses happened at two different points in space. They were simultaneous as observed by Rion on the moving platform, but were not simultaneous from Riel's perspective on the ground. Both perspectives were correct. Simultaneity is relative.
- That simultaneity is relative is a consequence of the constancy of the speed of light and the use of light to judge the order of events.
- Note the two light pulses reached Rion at a single point in space. Therefore, according to both inertial frames, Rion received the two light signals simultaneously.
- Note that in Galilean relativity, the emission of light from each end would be simultaneous in both frames of reference. As observed by Riel, the light pulse from the back has a greater speed ($c + v$) and this makes up for the longer distance it has to travel. The light pulse from the front has a slower speed ($c - v$), but this is compensated by a shorter distance.

Key concepts

Principle of simultaneity: When two events that occur at two different points in space are simultaneous in one inertial frame, they cannot be simultaneous in another inertial frame of reference.

The Lorentz transformations

Revised

Lorentz transformations equations relate the position and time of an event in some frame S to the position and time in another frame S′ that is moving relative to frame S.

The Lorentz factor

- The Lorentz transformation equations use the Lorentz factor, which is the factor by which the length, time and mass change due to the relative motion between frames S and S′.
- The Lorentz factor is equal to $\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$, where v is the relative speed between the two inertial reference frames S and S′ and c is the speed of light.

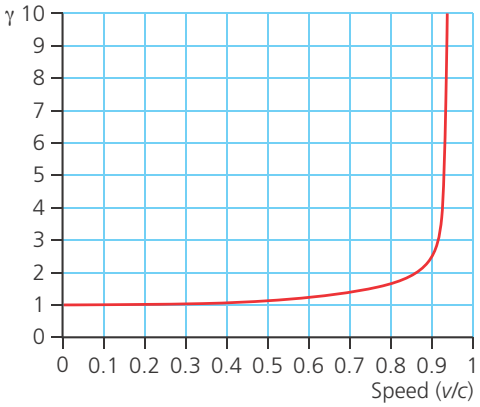


Figure 13.14

- Figure 13.14 shows a graph of the values of the Lorentz factor γ versus the ratio $\frac{v}{c}$ or the speed v as a fraction of the speed of light. It can be seen that the Lorentz factor is equal to 1 when $\frac{v}{c} \ll 1$, that is, when $v \ll c$ (low velocities). As the speed v gets closer to the speed of light, the value of the Lorentz factor increases rapidly.

QUESTIONS TO CHECK UNDERSTANDING

4 Calculate the Lorentz factor when $v = 0.1c$, $0.6c$.

Expert tip

It helps to remember some Lorentz factor values for some common velocities (Table 13.1). This can save you precious time during an examination.

Table 13.1 Lorentz factors for some common velocities

$\frac{v}{c}$	0.6	0.7	0.8	0.9	0.95
Lorentz factor	1.25	1.40	1.67	2.29	3.20

Transformation equations

Revised

- Consider two inertial frames, S(x, t) and S′(x', t'), that are moving relative to one another with velocity v. The coordinates (x, t) and space and time intervals (Δx, Δt) are related to coordinates (x', t') and to intervals ($\Delta x'$, $\Delta t'$), respectively, by the Lorentz transformation formulas, given in Table 13.2.

Table 13.2 Lorentz transformation formulas

	Lorentz transformation (given in the IB Physics data booklet)	Inverse transformation (not in the IB Physics data booklet)
Coordinate	$x' = \gamma(x - vt)$ $\Delta x' = \gamma(\Delta x - v\Delta t)$ $y = y'$ $z = z'$	$x = \gamma(x' + vt')$ $\Delta x = \gamma(x' + v\Delta t')$ $y' = y$ $z' = z$
Time	$t' = \gamma\left(t - \frac{vx}{c^2}\right)$ $\Delta t' = \gamma\left(\Delta t - \frac{v\Delta x}{c^2}\right)$	$t = \gamma\left(t' + \frac{vx'}{c^2}\right)$ $\Delta t = \gamma\left(\Delta t' + \frac{v\Delta x'}{c^2}\right)$
Notes	The Lorentz coordinate transformation is linear in x and x', hence a single event in S corresponds to a single event in S′. The Lorentz time transformation, t' is dependent on both time t and position x. Likewise, in the inverse transformation, t is dependent on t' and x'. This is unlike the Galilean transformation where $t = t'$. Both the coordinate and time transformations reduce to the Galilean transformation when $v << c$.	Note that to obtain the inverse transformation, the primed and unprimed quantities are exchanged ($x \leftrightarrow x'$, $u \leftrightarrow u'$) and the velocity v is replaced by $-v$ ($v \leftrightarrow -v$). This is consistent, since if S′ has velocity v in S, then S has velocity $-v$ in S′; since both are equally valid inertial reference frames.

- The Galilean coordinate, time and velocity transformations are incorrect because they do not keep the consistency of the laws of physics in the same form for the different inertial frames. The Lorentz transformations are the correct coordinate and time transformations as they preserve the invariance of all laws of physics in different inertial frames of reference.

Lorentz transformation of space, distance and time, and time interval measurements in different inertial frames of reference

Worked example

1 As measured in the laboratory, the velocity of an electron is 0.8c. It was also observed in the laboratory that at $t = 9.7 \times 10^{-9}$ s, it is at a position $x = 2.6$ m down the length of a vacuum tube.

a Calculate the value of the Lorentz factor.

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{1}{\sqrt{1 - \frac{0.64c^2}{c^2}}} = 1.67$$

b Determine the time and position of the electron according to an observer in the electron's reference frame.

$$x' = \gamma(x - vt) = 1.67\,(2.6 - (0.8 \times 3.0 \times 10^8) \times 9.7 \times 10^{-9}) = 0.45\,\text{m}$$

$$t' = \gamma\left(t - \frac{vx}{c^2}\right) = 1.67\left(9.7 \times 10^{-9} - \frac{0.8c \times 2.6}{c^2}\right) = 4.6\,\mu s$$

2 Two inertial observers are travelling with a relative velocity of 0.6c and both see two events occur. According to the observer in frame S′, the events occur 4.2 m apart and with a time interval of 2.4×10^{-8} s between them. According to the observer in S, what are the spatial (Δx) and time (Δt) intervals between the two events?

$$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} = \frac{1}{\sqrt{1-\frac{0.36c^2}{c^2}}} = 1.25$$

$$\Delta x = \gamma(\Delta x' + v\Delta t') = 1.25\,(4.2 + [0.6c \times 2.4 \times 10^{-8}]) = 10.7\,\text{m}$$

$$\Delta t = \gamma\left(\Delta t' + \frac{v\Delta x'}{c^2}\right) = 1.25\left[2.4 \times 10^{-8} + \frac{0.6c \times 4.2}{c^2}\right] = 40.5\,\text{ns}$$

3 A rocket of length 1200 m is moving at a speed 0.80*c* relative to the Earth. A light pulse is emitted from the back of the rocket and is received at the front of the rocket. Let E be the reference frame that is at rest with the Earth and R be the reference frame that is at rest with the rocket.

Figure 13.15 shows the situation from the view of E.

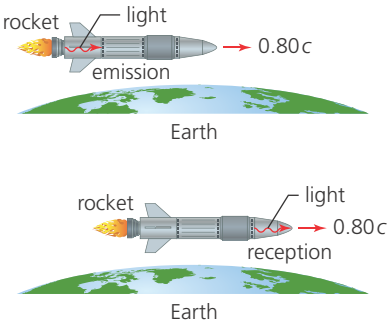


Figure 13.15

a Calculate the time taken between the emission and reception of the light pulse according to an observer in R and observer in E.

The distance travelled by the light pulse from emission to reception as seen from each frame is in Figure 13.16.

The time taken as seen from R:

$$\Delta t' = \frac{s}{v} = \frac{1200}{3.0 \times 10^8} = 4.0 \times 10^{-6}\,\text{s}$$

The time taken as seen from E:

$\gamma = 1.67$

At $v = 0.80c$,

$$\Delta t = \gamma\left(\Delta t' + \frac{v\Delta x'}{c^2}\right) = 1.67\left(4.0 \times 10^{-6} + \frac{0.80c \times 1200}{c^2}\right) = 1.20 \times 10^{-5}\,\text{s}$$

b Calculate the distance separating the emission and reception of the light pulse according to an observer in E.

$$\Delta x = \gamma(\Delta x' + v\Delta t') = 1.67(1200 + (0.80c \times 4.0 \times 10^{-6})) = 3600\,\text{m}$$

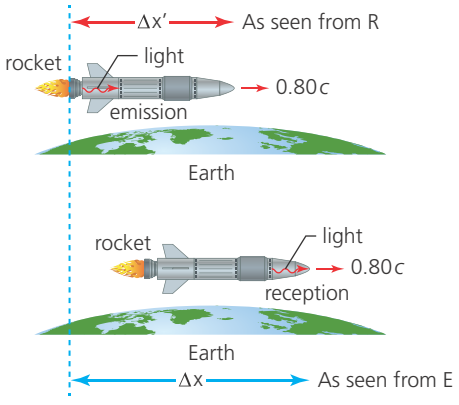


Figure 13.16

Lorentz transformation and simultaneity

- The principle of simultaneity can be proven by using the Lorentz time transformation equation which is equal to $\Delta t' = \gamma\left(\Delta t - \frac{v\Delta x}{c^2}\right)$.
- If two events happened simultaneously in frame S, then $\Delta t = 0$. If these two events happen at two different points in space in frame S, then $\Delta x \neq 0$. Then the Lorentz time transformation equation is reduced to $\Delta t' = \gamma\left(-\frac{v\Delta x}{c^2}\right) \neq 0$.

  This shows that when two events are simultaneous in frame S $\Delta t = 0$ and happen at two different points $\Delta x \neq 0$, then the same events are not simultaneous in frame S′ ($\Delta t' = 0$).
- However, when two events are simultaneous in frame S $\Delta t = 0$, but happen at same point $\Delta x = 0$, then the same events are also simultaneous in frame S′ ($\Delta t' = 0$).

Velocity addition

Revised

- The Galilean method of adding velocities is incorrect because it assumes that time and space are absolute. By using Lorentz transformation, Einstein formulated a different relativistic formula.

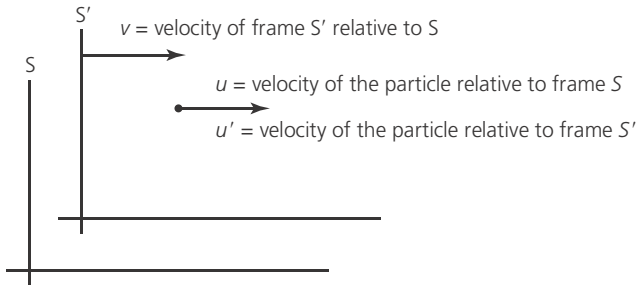


Figure 13.17

- Velocity addition formula: Consider Figure 13.17. Assume a particle is moving with a velocity u as measured in frame S. Another inertial frame S′ is moving with a velocity v with respect to frame S. Then the velocity u' of the particle as measured in frame S′ is given by:

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

- A closer look at the equation reveals the following:
  - If both u and v are very small compared to c, then $\frac{uv}{c^2} \to 0$, then the equation reduces to $u' = u - v$ which is the Galilean velocity addition formula.
  - If either $u = c$ or $v = c$ is, then $u' = c$.

Expert tip

To solve for u, the primed and unprimed quantities are exchanged ($u \leftrightarrow u'$) and the velocity v is replaced by $-v$ ($v \leftrightarrow -v$).

This is consistent, since if frame S′ has velocity v with respect to frame S, then frame S has velocity $-v$ with respect to frame S′; since both are equally valid inertial reference frames.

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

Solving problems involving velocity addition

Steps:

1. Draw a picture of the situation.
2. Select frames S, S′ and, hence, the velocities v, u and u'. Usually, the velocities of two objects are measured with respect to frame S. One of the moving objects is frame S′.
3. Substitute v, u and u' into the equation. Ensure that the algebraic signs of the velocities are consistent with the directions.
4. Understand the results.

Case 1 Two objects moving in the same direction

Two space shuttles approach planet X from the same direction. As observed from the planet, space shuttle A has a velocity of 0.75c, while space shuttle B has a velocity of 0.95c. Calculate the velocity of A as seen by B.

Draw the situation.

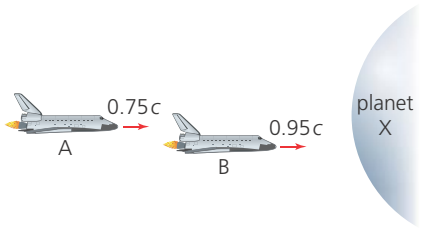


Figure 13.18

Select the frames and velocities:

Frame S is planet X since the velocities of the two rockets are measured with respect to the planet.

It follows that the space shuttle B is frame S′ since the problem requires the velocity of space shuttle A with respect to B.

Finally, $v = 0.95c$, $u = 0.75c$, u' is the velocity of A as seen from B.

Substitute into the equation.

A diagram of the situation in terms of frames S and S′ is shown below.

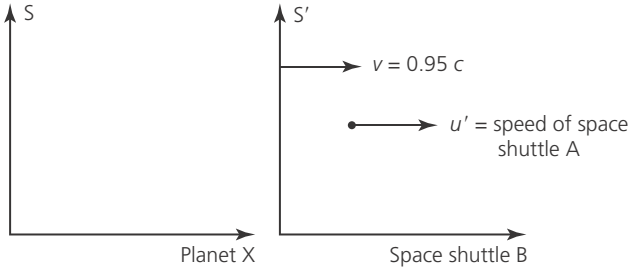


Figure 13.19

$$u' = \frac{u - v}{1 - \frac{uv}{c^2}} = \frac{0.75c - 0.95c}{1 - \frac{0.75c \times 0.95c}{c^2}} = -0.70c$$

Interpretation of the answer. The negative answer implies that as seen from space shuttle B, space shuttle is A moving to the left at a speed of 0.70c or that A is moving away from B.

A rocket ship is approaching a space station at a speed of 0.6c. It emits a light pulse towards the space station. Determine the speed at which an observer in the space station will see the light pulse arrive.

Draw the situation.

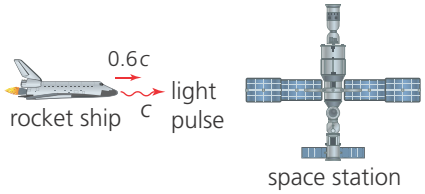


Figure 13.20

Select the frames and velocities:

Frame S: space station

Frame S′: rocket ship

$v = 0.6c$, $u' = c$, $u = ?$

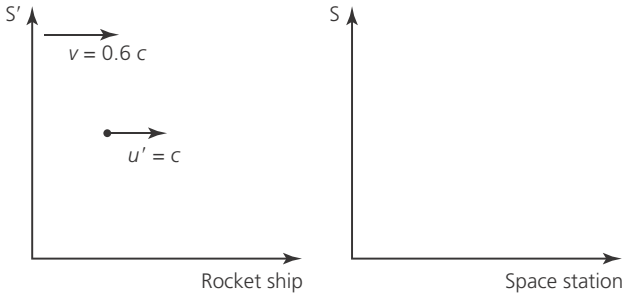


Figure 13.21

Since, u is the unknown, the inverse transformation of the velocity addition formula is necessary.

$$u = \frac{u' + v}{1 + \frac{u'v}{c^2}} = \frac{c + 0.6c}{1 + \frac{c \times 0.6c}{c^2}} = \frac{1.6c}{1.6} = c$$

Note that this result is consistent with the second postulate which states that the speed of light is constant.

Case 2 Objects moving in opposite directions

Two particles are seen from the laboratory frame to be travelling away from each other. Particle A moves to the right at 0.65*c*, while particle B moves to the left at 0.95*c*. Calculate the speed of particle B as seen by particle A.

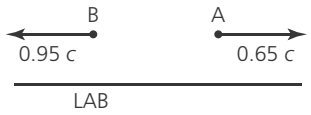


Figure 13.22

Frame S: Laboratory

Frame S′: Particle A

$v = 0.65c,\ u = -0.95\ c,\ u' = ?$

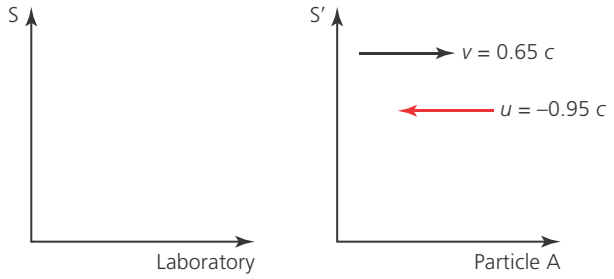


Figure 13.23

As seen by particle A, particle B is moving to the left. Also, since they are moving in opposite directions, particle B will appear to be faster.

$$u' = \frac{u - v}{1 - \frac{uv}{c^2}} = \frac{-0.95c - 0.65c}{1 - \frac{-0.95c \times 0.65c}{c^2}} = -0.99c$$

A rocket is coming from the right of a planet at a speed of 0.80*c*. A light pulse is approaching the planet from the left side. Show that the speed of the light pulse as seen from the rocket is c.

Figure 13.24

Frame S: Planet

Frame S′: Rocket

$v = -0.80c,\ u = c,\ u' = ?$

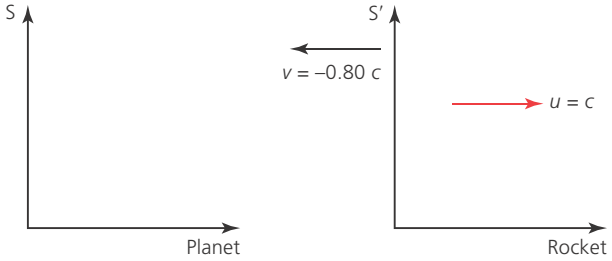


Figure 13.25

$$u' = \frac{u - v}{1 - \frac{uv}{c^2}} = \frac{c + 0.80c}{1 - \frac{c \times (-0.80c)}{c^2}} = c$$

Expert tip

The speed of separation between the two particles is equal to 0.65*c* + 0.95*c* = 1.60*c*. This is not a violation of the limit of the speed of light because the speed of separation is not a speed measured relative to an observer. Neither particle has a speed greater than *c* with respect to any observer (see www.uwgb.edu/fenclh/problems/modern/relativity/3/).

Invariant quantities

Revised

- In special relativity, a quantity is invariant when it remains unchanged under Lorentz transformation and thus it has the same value in all inertial frames. An invariant quantity is useful because it gives different observers a quantity about which they can all agree.
- Invariant quantities include space-time interval, proper time, proper length and rest mass.

Space-time interval

- Space-time is a four-dimensional continuum where time and three-dimensional space are merged together. Any point in space-time is given by the coordinates P(x, y, z, ct).
- In any inertial frame S, the space-time interval Δs is defined as $\Delta s^2 = (ct)^2 - (\Delta x^2 + \Delta y^2 + \Delta z^2)$.

- If motion is one-dimensional and along the x-axis, then $\Delta y = \Delta z = 0$ and thus, $\Delta s^2 = (c\Delta t)^2 - (\Delta x^2)$. Consequently, the space-time interval in inertial frame S′ that is moving relative to frame S is $(\Delta s')^2 = (c\Delta t')^2 - (\Delta x')^2$.
- Space-time interval between two events is the same in every inertial frame, i.e. invariant. In the Data Booklet, the space-time invariance is expressed as $(ct')^2 - (x')^2 = (ct)^2 - (x)^2$.

The Lorentz transformation equations can be used to prove space-time invariance.

Proper time

- **Proper time** is the time measured in an inertial frame where the events occur in the same point in space. In the reference frame where proper time is measured, $\Delta x = 0$.
- Proper time interval is always the shortest time interval. This is easily proven by using space-time invariance.

  Consider two inertial frames S and S′, moving relative to each other. The space-time interval invariance between the two frames is:

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

  Assume that the proper time occurs at frame S, thus proper time is Δt and $\Delta x = 0$. The above equation is reduced to:

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

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

  Therefore,

  $$(c\Delta t')^2 > (c\Delta t)^2$$

  $$\Delta t' > \Delta t$$

- The invariance of proper time can be proven using Lorentz transformation.
- Consider two inertial frames S and S′, moving relative to each other. The space-time interval invariance between the two frames is:

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

  Assume that the proper time occurs at frame S, thus proper time is Δt and $\Delta x = 0$. The above equation is reduced to:

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

  Using Lorentz transformation, $\Delta x' = \gamma(\Delta x - v\Delta t)$ and $\Delta t' = \gamma\left(\Delta t - \frac{v\Delta x}{c^2}\right)$, but $\Delta x = 0$.

  Thus

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

  $$(\gamma v\Delta t)^2 - (c\gamma\Delta t)^2 = -(c\Delta t)^2$$

  $$(\Delta t)^2\gamma^2(v^2 - c^2) = -(c\Delta t)^2$$

  $$(\Delta t)^2\left(-\frac{c^2}{v^2 - c^2}\right)(v^2 - c^2) = -(c\Delta t)^2$$

  $$(c\Delta t)^2 = (c\Delta t)^2$$

QUESTION TO CHECK UNDERSTANDING

5 A pair of events have space-time coordinates (12×10^9 m, 60 s) and (3×10^9 m, 10 s) in reference frame S′. Determine the proper time interval Δt measured in reference frame S.

Proper length

- **Proper length** is the length of an object measured in the reference frame where it is at rest.
- The length of an object is always longest when measured in the reference frame where it is at rest.

- One prediction of the special theory of relativity is length contraction, that is, the length of an object becomes shorter when measured in a moving reference frame. This will be discussed in more detail later.

Rest mass

- **Rest mass** is the mass of the body measured in the reference frame where it is at rest.
- Special relativity has shown that inertial mass is not constant and increases with velocity.
- Closely related to the rest mass is another invariant quantity, the **rest energy**, $E_0 = m_0c^2$. The equation implies that mass is a concentrated form of energy; that mass can be converted into energy or energy into mass.
- Rest energy is real. Nuclear fission has shown that energy can be obtained from mass defects that occur during the process. Elementary particle interactions have shown that high energy collisions result in particle creation.

Time dilation and length contraction

Revised

Time dilation

- In the special theory of relativity, **time dilation** is a phenomenon where time in a moving inertial frame is observed to run slow or dilated.
- Time dilation is dependent on the relative speed between the different inertial frames of reference. The faster the speed is, the greater is the time dilation.
- The time dilation effect is given by the formula:

$$\Delta t = \gamma \Delta t_0$$

where γ is the Lorentz factor and Δt_0 is the proper time.

- Note that due to the effect of the Lorentz factor in the formula, time dilation is negligible at low speeds.
- Time dilation is real. Even periodic events slow down accordingly, such as radioactive decay or biological processes like heartbeat, metabolism and aging. Note that all these clocks work under a complicated combination of laws of mechanics and electromagnetism. Time dilation is not an illusion.
- Time dilation is symmetric.
  - Consider Figure 13.26 which shows the times in frames S and S′, where frame S′ is moving relative to frame S. An observer in frame S will see that time in frame S′ runs slow, e.g. 4.0 s have gone past his clock, but only 3.5 s in frame S′.
  - However, an observer in frame S′ will see that time in frame S is running slow. He observes a time interval of 4.0 s in his frame, but only 3.5 s have gone in frame S.
  - Who is correct? Both are! This is because the situation is symmetrical, i.e. both could say that he is at rest and the other is moving. From the first postulate, motion is relative and there is no preferred frame of reference.

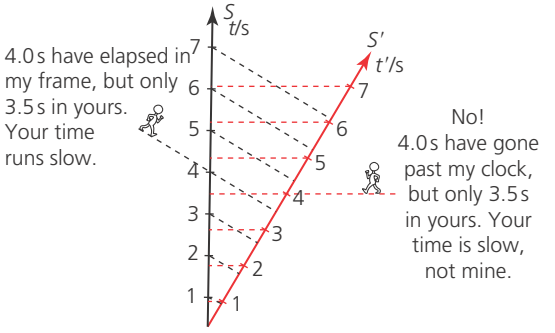


Figure 13.26 Time dilation is symmetric.

Length contraction

- Length contraction is proportional to the object's speed relative to the observer.
- Length contraction applies only to the dimension that is parallel to the direction of the velocity, as shown in Figure 13.27. The other dimensions are not affected by length contraction.

Key concept

Length contraction is a relativistic effect where a decrease in the length is detected by an observer in objects or distances that are moving relative to him.

- The formula for length contraction is given by:

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

where L = relative length; L_0 = proper length; and γ = Lorentz factor.

Solving problems involving time dilation and length contraction

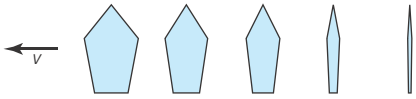


Figure 13.27 Length contraction. This is what a pentagon might look like as it approaches relativistic speeds. Only the dimension parallel to the velocity contracts.

Worked example

The distance between the Earth and a distant star is 6.0×10^{16} m as measured by an observer on Earth. An astronaut in a rocket ship from the Earth travels towards the star at a speed of $0.8c$.

1 Calculate the time taken by the astronaut to travel from the Earth to the star as measured

a by an Earth observer:

As measured by an Earth observer the distance to the star is 6.0×10^{16} m.

$$\text{Thus } t = \frac{s}{v} = \frac{6.0 \times 10^{16}}{0.8c} = \frac{6.0 \times 10^{16}}{0.8 \times 3.0 \times 10^{8}} = 2.5 \times 10^{8}\,\text{s}$$

b by the astronaut:

When $v = 0.8c$, $\gamma = 1.67$.

As measured by the astronaut, the distance to the star is contracted.

$$L = \frac{L_0}{\gamma} = \frac{6.0 \times 10^{16}}{1.67} = 3.59 \times 10^{16}\,\text{m}$$

$$\text{Therefore, } t = \frac{s}{v} = \frac{3.59 \times 10^{16}}{0.8c} = \frac{3.59 \times 10^{16}}{0.8 \times 3.0 \times 10^{8}} = 1.50 \times 10^{8}\,\text{s}$$

2 State and explain which of the times obtained in (a) above, measured by the Earth observer or by the astronaut, is the proper time.

The time measured by the astronaut is the proper time. The two events, leaving the Earth and arriving at the star, happened at the same point in the reference frame of the astronaut. (Besides, the values should give you a hint. Proper time is always the shorter time.)

Derivation of the time dilation equation using the Lorentz transformation

The time dilation formula can be derived from Lorentz transformation using the following steps.

Steps	Details of each step
Define inertial frame where proper time occurs.	Consider two events that happen at the same point in S′ and thus are separated by a proper time interval $\Delta t_0 = \Delta t' = t'_B - t'_A$.
Solve time interval in the other frame by using inverse Lorentz transformation.	The time interval measured in S is: $t_B = \gamma\left(t'_B + \frac{vx'_B}{c^2}\right) t_A = \gamma\left(t'_A + \frac{vx'_A}{c^2}\right)$ $t_B - t_A = \gamma\left[(t'_B - t'_A) - \frac{v}{c^2}(x'_B - x'_A)\right]$
Apply condition for time dilation: $\Delta x = 0$	Since the two events happen at the same point in S′, then $x'_B - x'_A = 0$. $t_B - t_A = \gamma(t'_B - t'_A)$ $\Delta t = \gamma \Delta t_0$ (time dilation formula)

Derivation of the length contraction equation using the Lorentz transformation

The length contraction formula can be derived from the Lorentz transformation using the following steps:

Steps	Details of each step
Define inertial frame where proper length is measured.	Consider a rod at rest in inertial frame S′. The endpoints of the rod are x'_A and x'_B, thus its length $L_0 = x'_B - x'_A$, is the rest length or proper length of the rod.
Solve for length in the other frame.	Using Lorentz transformation, the proper length $L_0 = x'_B - x'_A$ is equal to: $L_0 = x'_B - x'_A = \gamma(x_B - vt_B) - \gamma(x_A - vt_A)$ $L_0 = x'_B - x'_A = \gamma[(x_B - x_A) - v(t_B - t_A)]$
Apply condition for length measurement: $\Delta t = 0$	In order to measure the length of the rod, the coordinates of the endpoints must be determined at the same time, that is $t_A = t_B$ or $(t_B - t_A) = 0$. Hence the above equation reduces to: $L_0 = \gamma(x_B - x_A)$ $L_0 = \gamma L$ where $L = x_B - x_A$, the relative length of the rod.

Expert tips

Applying the Lorentz transformation, time dilation and length contraction equations

When do you use Lorentz coordinate transformation against the length contraction formula?

- The Lorentz coordinate transformation applies to distances that change with time. For example, the distance between two objects moving relative to each other.
- The length contraction formula applies to distance that is independent of time. Examples are the distance between two stationary objects or the length of a ruler.

When do you use Lorentz time transformation or the time dilation formula?

- The Lorentz time transformation applies to events that do *not* happen at the same point in space in any frame of reference.
- The time dilation formula applies to only events that happen in the same point in space in a frame of reference.

However, when in doubt, apply the Lorentz transformation. As shown previously, the Lorentz transformation equations reduce to the time dilation or length contraction formula.

The muon decay experiment

Revised

- Muons are particles that are created when cosmic rays from deep space collide with nuclei of atoms in the Earth's upper atmosphere. They have the same properties as electrons except that they are around 200 times more massive. Muons travel toward the Earth's surface at a speed of 0.995*c*.
- In 1963, Smith and Frisch conducted an experiment to test time dilation. Muon detectors were placed on top of Mount Washington, which was around

2 km high, and others placed at sea level. Smith and Frisch compared the number of muons detected from the two locations. The experiment is depicted schematically in Figure 13.28.

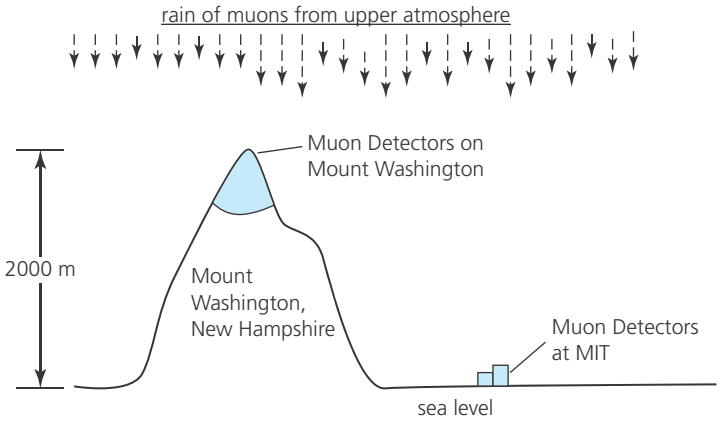


Figure 13.28 The muon decay experiment. Differences in the count rate between the detectors in the two locations verified the special theory of relativity.

- Muons are highly unstable and have a very short half-life of 1.5×10^{-6} s as measured in the muon's reference frame. The mean count rate at the top of the mountain from the experiment was around 570 muons per hour. From the mountain top, the muons were not expected to live long enough to reach the surface at sea level. The expected count rate at sea level was 36 muons per hour. However, a count rate of around 412 muons per hour was recorded. It was around 11 times more than was expected.
- This result can be explained through time dilation.
  - The half-life of 1.5×10^{-6} s is the proper time and is measured from the frame of reference of the muon.
  - At $v = 0.995c$, the Lorentz constant $\gamma = 10$. As measured by observers on Earth, the half-life of the muon was 10 times longer, i.e. time ran slower in the muon frame. Thus, the muons have a longer time to exist before they undergo decay.
- Special relativity predicted a count rate of around 428 muons per hour at sea level – in agreement with the measurements obtained by Smith and Frisch.
- The experiment verified that time dilation is real.

Solving problems involving muon decay experiment

Muons are particles that are created when cosmic rays from deep space collide with atoms in the Earth's upper atmosphere, around 15 km above the Earth's surface. Muons have a short half-life of 1.5×10^{-6} s as measured in their own frame of reference, and travel at $0.995c$.

1 Explain, with calculations, why it is not expected to detect muons on the Earth's surface.

Using Newtonian mechanics, the time the muons would travel to reach the ground is $t = \frac{s}{v} = \frac{15000\,\text{m}}{0.995c} = 5.0 \times 10^{-5}\,\text{s}$.

The number of half-lives the muons go through during this time is $\frac{5.0 \times 10^{-5}}{1.5 \times 10^{-6}} = 31$.

With only a tiny fraction remaining $\left(\frac{1}{2}\right)^{31} = 4.7 \times 10^{-10}$, they are not expected to be detected on the Earth's surface.

2 However, a large fraction of the muons is observed on the ground. Outline how this can be explained using time dilation or length contraction.

When $v = 0.995c$, the Lorentz factor $\gamma = 10$.

Time dilation explanation: From the Earth's frame, due to time dilation, the half-life of the muon was $nt = \gamma nt_0 = 10 \times 1.5 \times 10^{-6}\,s = 1.5 \times 10^{-5}\,s$.

Therefore, the number of half-lives the muons underwent before they reached the Earth was $\frac{5.0 \times 10^{-5}\,s}{1.5 \times 10^{-5}\,s} = 3$.

Thus, the muons only decayed through three half-lives before reaching the ground and a significant fraction remained.

Length contraction explanation

- From the muon's perspective, the altitude of the atmosphere (15 km) becomes contracted by a factor of $\gamma = 10$ to around 1.5 km, i.e. $L = \frac{L_0}{\gamma} = \frac{15\,km}{10} = 1.5\,km$.
- The travelling time of the muons was $t = \frac{s}{v} = \frac{1.5\,km}{0.995c} = 5.0 \times 10^{-6}\,s$.
- The number of half-lives the muons underwent before they reached the Earth was $\frac{5.0 \times 10^{-6}\,s}{1.5 \times 10^{-6}\,s} = 3$.

Thus, like in time dilation explanation, the muons would only decay through three half-lives before reaching the ground and a significant fraction remained.

Table 13.3 gives a summary of what each frame observes. The Earth frame sees time dilation while the muon frame observes length contraction. Note that the conclusions for both frames are consistent with each other.

Table 13.3

	Earth frame	Muon frame
Distance travelled	15 km	1.5 km
Half-life	15 μs	1.5 μs
No. of decays	3	3
Remaining fraction	$\frac{1}{8}$	$\frac{1}{8}$

13.3 Space-time diagrams

Revised

Essential idea: Space-time diagrams are a very clear and illustrative way to show graphically how different observers in relative motion to each other have measurements that differ from each other.

- A space-time diagram is a representation of space-time obeying the laws of special relativity. It is drawn similar to a rectangular coordinate system, but time is in the vertical axis and space is in the horizontal axis.
- Events are represented as dots in a space-time diagram.

Representing events in a space-time diagram

Revised

- Figure 13.29 shows three events in a space-time diagram.
  - Events A, B and C are represented as dots in space time. Event A has coordinates (x_1, t_1). Event B has coordinates (x_2, t_2) and event C has coordinates (x_1, t_2).
  - Events A and C happened at the same point in space but at different times. Events that occur at the same point in space form a line parallel to the *t*-axis.
  - Events B and C occurred at the same time but at different locations in space. Simultaneous events form lines that are parallel to the *x*-axis.
- The worldline shows the path of an object in space-time.

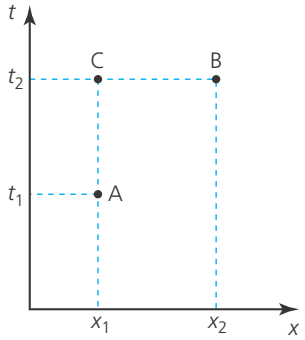


Figure 13.29

Representing the positions of a moving particle on a space-time diagram by a worldline

Revised

Figure 13.30 shows different worldlines.

- Line 1 represents a particle that is at rest. It stays at the same position as time progresses.
- Lines 2 and 3 represent particles that are moving to the right with a constant velocity. Line 3, that with a lower gradient, indicates a higher speed.
- Line 4 represents a particle that is accelerating.

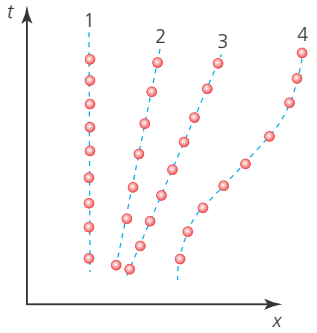


Figure 13.30

A more convenient space-time diagram uses *ct* in the vertical axis, instead of *t*. This makes both axes have the dimension of distance. The *ct* axis represents the time taken by light to travel one unit of distance. Figure 13.31 shows this type of space-time diagram.

- In this space-time diagram, the gradient of the worldline is equal to $\frac{c}{v}$, as derived below.

$$\text{gradient} = \frac{\Delta(ct)}{\Delta x} = \frac{c\,\Delta t}{\Delta x} = c \times \frac{\Delta t}{\Delta x} = c \times \frac{1}{v} = \frac{c}{v}$$

- Since the gradient is equal to $\frac{c}{v}$, the worldline of light ($v = c$) would have a gradient of 1, as shown in Figure 13.31a.

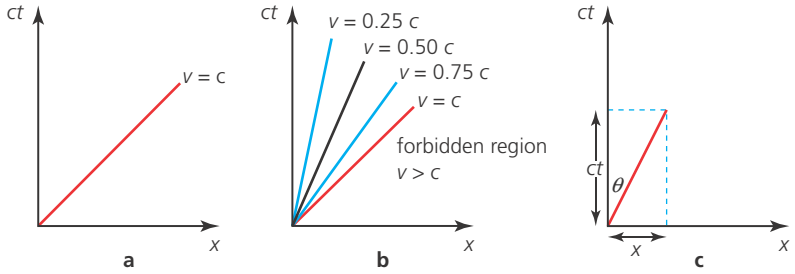


Figure 13.31 Space-time diagrams with a *ct* vertical axis.

- Figure 13.31b shows the worldlines of the particles travelling at different speeds. It can be seen that the faster the speed of the particle is, the less steep the line becomes. Since it is not possible for particles to exceed the speed of light, it is not possible to have a worldline below that of light.

- Figure 13.31c shows a worldline that makes an angle θ with the vertical. This angle is equal to $\theta = \tan^{-1}\left(\frac{v}{c}\right)$, as derived below.

$$\tan\theta = \frac{x}{ct} = \frac{1}{c} \times \frac{x}{t} = \frac{v}{c}$$

QUESTIONS TO CHECK UNDERSTANDING

6 A particle is moving at a constant speed of 0.8*c* to the right. Another particle is travelling at constant speed of 0.6*c* to the left. Draw the worldlines of the two particles in a space-time diagram.

7 A flashbulb at the centre of a train ($x = 0$) emits pulses that travelled towards opposite ends of the train. Draw the worldlines of the two pulses.

8 Four flashes, A, B, C and D, occurred at different points in space-time as shown in Figure 13.32 below. Determine the order at which the pulses occurred in each frame of reference.

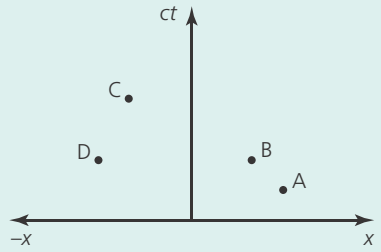


Figure 13.32

Representing more than one inertial frame on the same space-time diagram

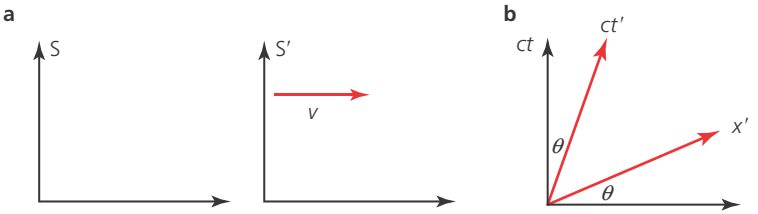


Figure 13.33

Consider frame S′ which has a rightward velocity v relative to frame S. Figure 13.33 shows the representation of the two inertial frames in a space-time diagram.

- The *ct*′-axis makes an angle $\theta = \tan^{-1}\left(\frac{v}{c}\right)$ with the *ct*-axis.
- The *x*′-axis is not drawn perpendicular to the *ct*′-axis. Instead, the *x*′-axis makes an angle $\theta = \tan^{-1}\left(\frac{v}{c}\right)$ from the *x*-axis.

Figure 13.34 shows how to obtain the coordinates of different events.

- To find the position coordinates of the events, a line parallel to the *ct*′-axis is drawn from the event to the *x*′-axis, as shown by the red lines. To find the time coordinates, a line parallel to the *x*′-axis is drawn from the event to the *ct*′-axis, as shown by the blue lines.
- The lines of simultaneity are drawn parallel to the *x*′-axis. Note that events B and C are simultaneous in frame S′.

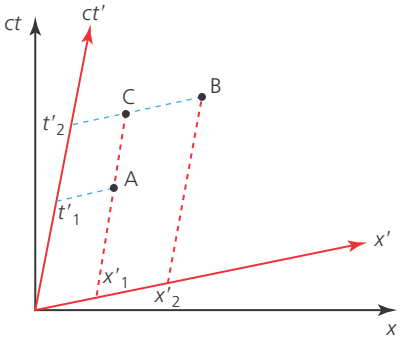


Figure 13.34

Simultaneity and space-time diagrams

Figure 13.35 below shows three events, A, B and C, in a space-time diagram.

- In frame S, it can be seen that events A and B, which occurred at two different points, were simultaneous. However, in frame S′, event B happened before event A. Thus, events A and B were simultaneous in frame S but not in frame S′.
- Likewise, events B and C are simultaneous in frame S′ but not in frame S.

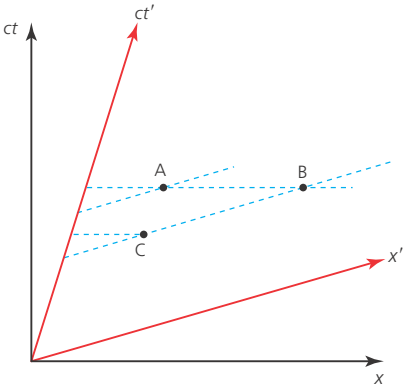


Figure 13.35

Worked example

The diagram below shows a train travelling to the right at constant velocity. A flashbulb is hanging from the ceiling of the train midway between the ends L and R of the train. Each flash sends single pulses in opposite directions.

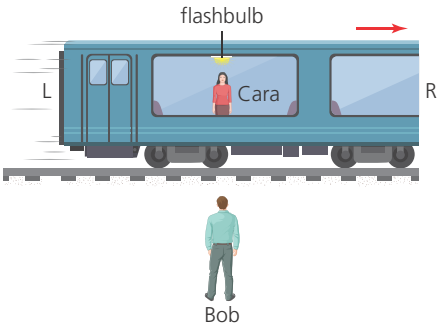


Figure 13.36

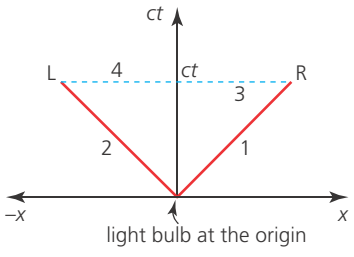
Cara is at rest at the centre of the train. Light pulses from the flashbulb were seen by Cara to hit the opposite walls L and R of the train simultaneously. Bob is at rest on the ground. He is opposite Cara at the moment the bulb flashed. The situation is shown in Figure 13.36.

Draw a space-time diagram to deduce whether or not the events are simultaneous to Bob.

Choose the frame S:

- The moving train's frame is convenient to use as frame S since the two events are simultaneous in Cara's frame. The events are represented by Figure 13.37a.
- Since light travels in opposite directions, we can place the light bulb at the origin. Lines 1 and 2 represent the two light pulses reaching ends L and R simultaneously. The time they reach ends L and R is projected to the *ct*-axis by lines 3 and 4. Note that these lines are parallel to the *x*-axis.

a



b

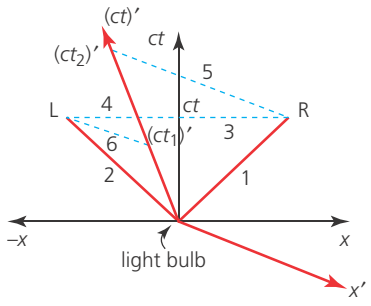


Figure 13.37

- Draw frame S′.
- From Cara's perspective, Bob is travelling to the left. Therefore, frame S′ is drawn as shown in Figure 13.37b.
- The time the light pulses reached ends L and R as seen by Bob is projected to the *ct*′-axis by lines 5 and 6. These lines are parallel to the *x*′-axis.
- It can be seen that in frame S′, the pulse reached end L first. Thus, the two light pulses did not reach ends L and R simultaneously as seen by Bob.

Time dilation in space-time diagram

Revised

- Time dilation in a space-time diagram is shown in Figure 13.38.
- Consider frame S where two events occur in the same place and thus, measures the proper time. Assume that the time interval between these two events in frame S is equal to $c\Delta t$, as shown in the ct-axis.
- The time measured in frame S′ can be obtained by transforming $c\Delta t$ on to the ct'-axis, as shown by line A. The time interval for the two events in frame S′ is equal to $c\Delta t'$.
- It can be seen from the diagram that the length of $c\Delta t'$ is greater than the length of $c\Delta t$. Therefore, $\Delta t' > \Delta t$. This means that observers in S′ see time run slower than in frame S.

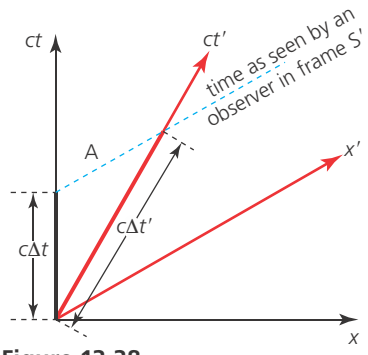


Figure 13.38

Length contraction in space-time diagram

Revised

- In any inertial frame, the length of the object is equal to the distance between two points that is measured simultaneously. This implies that all length measurements should use the line of simultaneity.
- Length contraction in a space-time diagram is shown in Figure 13.39. Again, frame S′ is moving relative to frame S.
- A rod is at rest relative to frame S′ and has a proper length L_0. This means that in frame S′, when an observer measures the distance between the front and the back of the rod simultaneously, he obtains the measurement as L_0.
- When an observer in frame S measures the distance between the front and the back of the moving rod simultaneously, he would obtain the length as L. The length in frame S has to be measured using a horizontal line of simultaneity parallel to the x-axis. It is easy to see that the length L is shorter than the proper length L_0. This is length contraction.
- Using the invariance of space-time interval, it is easy to see that $L^2 = L_0^2 - (c\Delta t')^2$. This further proves that $L < L_0$. Note that Euclidean geometry (i.e. Pythagorean theory) does not apply here.

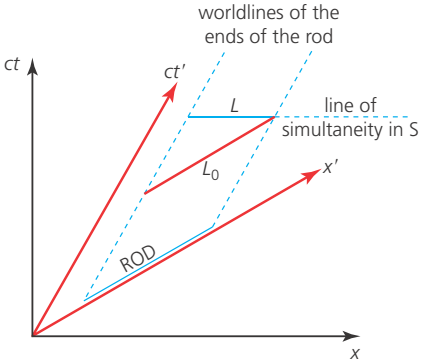


Figure 13.39

The twin paradox

- A paradox is a situation where contradictory conclusions are reached using valid deductions from acceptable premises. If the paradox cannot be resolved, then the theory in question failed and it should either be modified or rejected.
- The **twin paradox** involves a set of hypothetical twins, one of whom stayed on Earth (i.e. Earth twin) while the other took a relativistic round trip journey to and from space (i.e. travelling twin). Since both twins moved with respect to each other, then each twin observed time dilation in the other twin's reference frame. The paradox arises because each twin observed the other twin as travelling and himself as being at rest. Thus, each believed the other twin's clock ran slower and should be younger when they meet again.
- To resolve the paradox, one must recognize that since the travelling twin made an outward and return journey, the situation was no longer symmetrical. The Earth twin stayed in a single inertial frame, but the travelling twin did not. Disregarding the accelerating and decelerating frames during the turnaround, at the very least the travelling twin was in two different inertial frames – one when leaving the Earth and another when returning to the Earth. Only the travelling twin felt the acceleration at the turnaround point. This means that he cannot be regarded as being at rest and the Earth twin as being in motion.

- Since only the Earth twin remained in a single inertial frame, only he can fully apply the time dilation to the other reference frame. As observed by the Earth twin, time moves slower in the frame of reference of the travelling twin. Thus, the travelling twin aged slower.
- Also, due to length contraction, the Earth twin would measure the travelling twin's journey to be longer than that measured by the travelling twin. Since both twins measured the same relative speed, the longer distance measured in the Earth frame would correspond to a greater time interval. That is, time moves slower in the travelling twin's reference frame.

- In summary, time dilation is symmetric only when the two inertial frames are at constant relative motion. In the twin paradox problem, there are three inertial frames instead of two and this makes the situation non-symmetrical.

Space-time resolution of the twin paradox

- Figure 13.40 shows the space-time diagram of the twin paradox drawn from the perspective of the Earth twin.

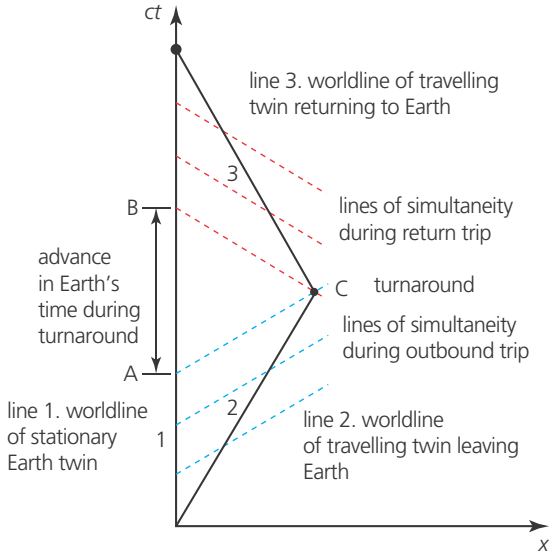


Figure 13.40

- Line 1 in the *ct*-axis shows the worldline of the Earth twin. He was stationary in the space-axis but moved forward in time.
- Line 2 is the worldline of the travelling twin when he left the Earth. Line 3 is the worldline of the travelling twin when he returned to Earth. The turnaround happened at C.
- The blue lines are lines of simultaneity for the travelling twin during the outbound part of the journey. The red lines represent the lines of simultaneity during the return or inbound leg of the journey.
- During the outbound segment (line 2) of the journey, time dilation was symmetrical. Each twin observed the other clock to be running slower. The symmetry is broken at point C when the travelling twin made the turnaround, that is, when he jumped to a different inertial frame.
- Note that during the jump, the line of simultaneity changed from blue to red. The relative simultaneity between the new inertial frame and the Earth frame had changed and the travelling twin saw that the time on Earth has suddenly advanced from A to B. This implied a sudden increase in the age of the Earth twin. Due to this effect, the travelling twin would be the younger twin when they meet again.

13.4 Relativistic mechanics

Revised

Essential idea: The relativity of space and time requires new definitions for energy and momentum in order to preserve the conserved nature of these laws.

Total energy and rest energy

Revised

- Rest energy is the energy of a particle at rest due to its rest mass m_0. This is equal to $E_0 = m_0c^2$.
- In relativistic mechanics the total energy is defined as $E = \gamma m_0c^2$ where γ is the Lorentz factor. This equation shows that the mass of a moving object increases by a factor of γ.
- The graph in Figure 13.41 shows how the mass increases as it approaches the speed of light. Note that relativistic increase in the mass becomes noticeable at around 0.5c.

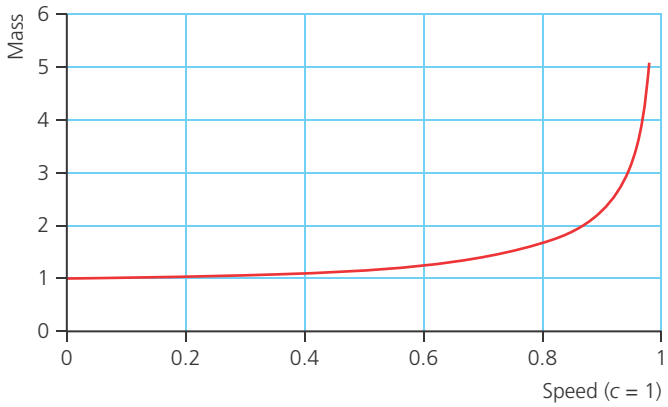


Figure 13.41

- Note that the total energy E of a moving particle is equal to the sum of the kinetic energy E_K and rest energy of the particle E_0.

  $E = E_0 + E_K$
- Conservation of energy. In relativistic interactions like high speed collisions of particles, it is the total energy $E = \gamma m_0c^2$ that is conserved. This suggests that it is possible for particles to be created from energy or energy released from the annihilation of particles.
- From the equations $E = \gamma m_0c^2$ and $E_K = E - E_0$, an expression for the kinetic energy is obtained as $E_K = (\gamma - 1)m_0c^2$.

Relativistic momentum

Revised

- Relativistic momentum is defined so that it remains conserved under Lorentz transformation.
- For a particle of rest mass, m_0, and with a velocity, v, measured in the observer's inertial frame, its relativistic momentum is defined as $p = \gamma m_0 v$.
- The velocity v in the definition is the velocity of the object in the observer's inertial frame. Thus, for a particle of rest mass m_0 that moves at a speed u as measured in S and u' as measured in S′ where S and S′ have a relative velocity v, the expressions for the momentum are given below.

  The momentum of the particle as measured by an observer in S is:

$$p = \frac{m_0 u}{\sqrt{1 - \frac{u^2}{c^2}}}.$$

The momentum of the particle as measured by an observer in S′ is:

$$p' = \frac{m_0 u'}{\sqrt{1 - \frac{u'^2}{c^2}}}$$

- Note that at low velocities, $\gamma = 1$ and the relativistic definition of momentum reduces to the classical momentum $p = m_0 v$.

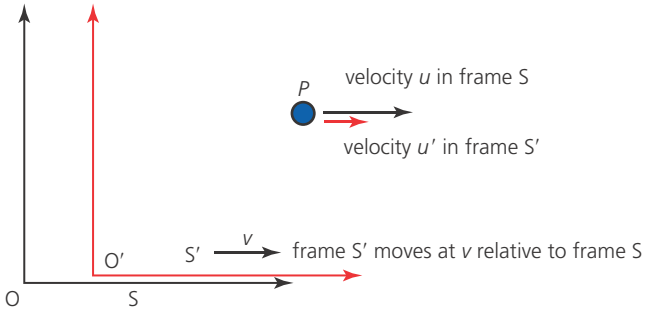


Figure 13.42

Relationship between energy and momentum

Revised

- The equation relating the total energy E, momentum p and rest energy $E_0 = m_0 c^2$ is:

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

- A plot of the relation between the total energy E and relativistic momentum p is shown in Figure 13.43. It can be seen that at low velocities (i.e. small p), the energy is increasing proportional to p^2 as Newtonian mechanics require. However, at relativistic speeds (i.e. large p), the energy becomes more and more nearly proportional to the momentum p, as special relativity requires.

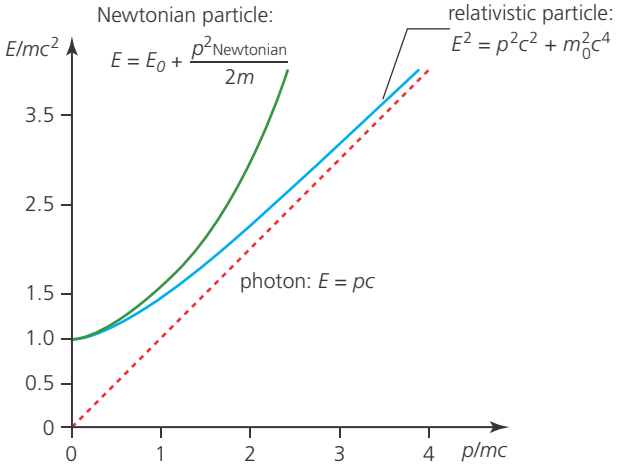


Figure 13.43

- The above equation can be rewritten as $E^2 - p^2 c^2 = m_0^2 c^4$. Since the rest mass m_0 and c are invariant quantities, then the difference $E^2 - p^2 c^2$ must be the same for all inertial frames.

Photons

Revised

- Photons have zero rest mass ($m_0 = 0$) and thus, the energy-momentum equation becomes $E^2 - p^2 c^2 = 0$ and consequently, $E = pc$.
- Since $E^2 - p^2 c^2$ is invariant, then $E = pc$ must also be true in all inertial frames.

- Since the photon carries energy, it has momentum even when it is massless. As the energy of a photon is $E = hf$, then $hf = pc$. This is consistent with de Broglie's equation, $p = \frac{h}{\lambda}$.

Electric charge is an invariant quantity

Revised

- Electric charge does not depend on time or position and is the same in all inertial frames. This prompts Maxwell's equations to be invariant under Lorentz transformation.
- Invariance of electric charge is also consistent with the law of conservation of charges – that electric charges cannot be created nor destroyed, but can only be transferred from one location to another.

QUESTION TO CHECK UNDERSTANDING

9 A nuclear particle has a rest mass of 6.0 GeV c^{-2} and momentum of 8.0 GeV c^{-1}. Calculate the particle's energy and speed.

Particle acceleration

Revised

- In Newtonian mechanics, when a particle of charge q is accelerated by a potential difference V, it gains kinetic energy equal to $\Delta E_K = qV$ or $\frac{1}{2}m_0v^2 = qV$. This equation implies that a very large potential difference V could lead to a very large speed v that can be greater than the speed of light. This is not possible under relativistic mechanics.
- To determine the energy gained by a charged particle q that is accelerated by a potential difference V, we use the definition of the total energy where $E_K = qV$.

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

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

QUESTION TO CHECK UNDERSTANDING

10 Electrons are accelerated from rest through a potential difference of 2.5 MV.

a Use Newtonian mechanics to determine the speed of the electrons in terms of c.

b Use relativistic mechanics to determine the speed of the electrons in terms of c.

Solving problems involving relativistic energy and momentum conservation in collisions and particle decays

QUESTIONS TO CHECK UNDERSTANDING

11 A proton collided head-on with another proton, each with the same total energy. The following reaction occurs:

$$p^+ + p^+ \rightarrow p^+ + p^+ + p^- + p^+$$

where p^+ is a proton and p^- is an anti-proton. They each have a rest energy of 935 MeV c^{-2}.

a Explain how it is possible that the total mass after collision is greater than the total mass before collision.

b Show that for this reaction to occur, the minimum total energy of each colliding proton is 1870 MeV. State any assumption you made.

c Determine the momentum of the proton whose total energy is 1870 MeV.

12 An electron with kinetic energy E_K = 1.000 MeV makes a head-on collision with a positron at rest. In the collision, the two particles annihilate each other and are replaced by two photons of equal energies. The reaction is:
$e^- + e^+ \rightarrow 2\gamma$

a Determine the energy of each photon.

b Calculate the

i momentum of each photon

ii wavelength of each photon.

13.5 General relativity

Revised

Essential idea: General relativity is applied to bring fundamental concepts of mass, space and time in order to describe the fate of the universe.

The equivalence principle

Revised

- The principle of equivalence states that observations made in an accelerated frame of reference are indistinguishable from observations made in a stationary frame of reference in a gravitational field. That is, the effects of acceleration and the effects of gravity are equivalent.
- The principle of equivalence can be illustrated from the thought experiment shown in Figure 13.44.
  - Figure 13.44a shows a frame of reference that is at rest inside the Earth's gravitational field. When a ball is thrown horizontally, an observer on Earth sees the object to move in a parabolic path (i.e. projectile motion).
  - Figure 13.44b shows a stationary reference frame in a gravity-free region. When the ball is thrown horizontally, in the absence of a net force on it, it will move in a straight horizontal line.
  - Figure 13.44c shows a reference frame rocket being accelerated upward in a gravity-free region. From the outside, the ball that is thrown horizontally is seen to travel in a straight path. However, for an inside observer, the ball is seen to follow a parabolic path as in a gravitational field. Thus, the acceleration produces a gravitational effect.

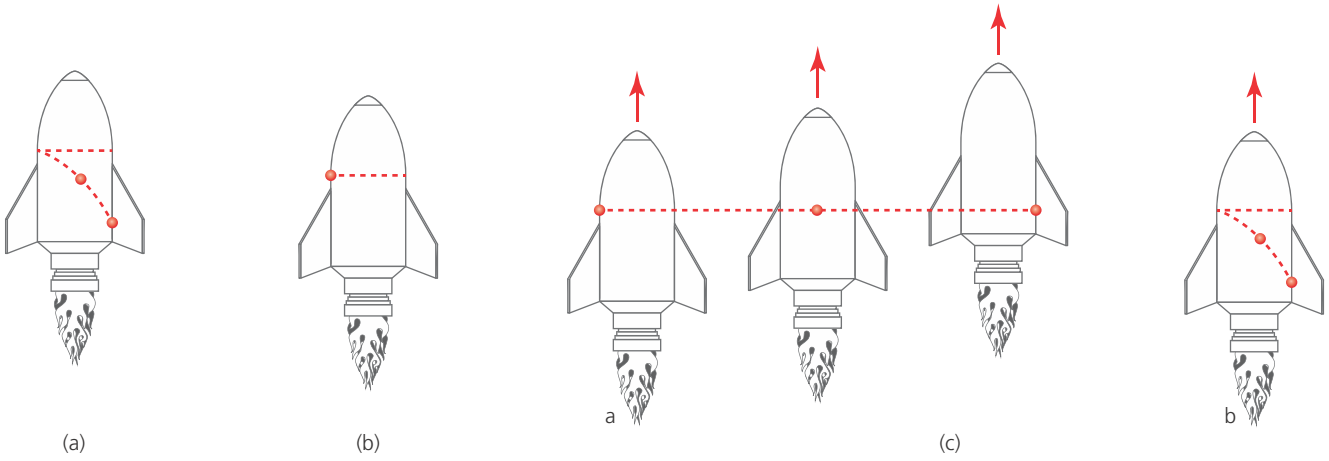


Figure 13.44 Thought experiments for principle of equivalence. (a) The rocket stationary on Earth. The gravitational field causes the parabolic motion of the ball. (b) The rocket at rest in a gravity-free region. The ball moves horizontally in a straight line. (c) The rocket in a gravity free region but accelerating upward. Inside the acceleration produces a gravitational effect.

- Einstein stated that this principle is true for all physical laws including mechanical, electromagnetic and optical phenomena.

Using the equivalence principle to deduce bending of light near massive objects

- The principle of equivalence predicted that a gravitational field can bend light.
- This can be illustrated from the thought experiment in Figure 13.45 in which there is a glass sheet in the centre of each rocket.
  - Figure 13.45a shows the rocket at rest. A light pulse projected horizontally will be seen to move in a straight horizontal path.
  - Figure 13.45b shows a rocket with a constant upward velocity. Outside the rocket (i), the light pulse is seen to travel in a straight horizontal path. Since the rocket moves equal distance increments every second, an inside observer (ii) will see the light to move straight but sloping downward.
  - Figure 13.45c shows a rocket with a constant upward acceleration. Since the rocket covers more distance every second, an inside observer (ii) sees the light pulse to move in a curved path.

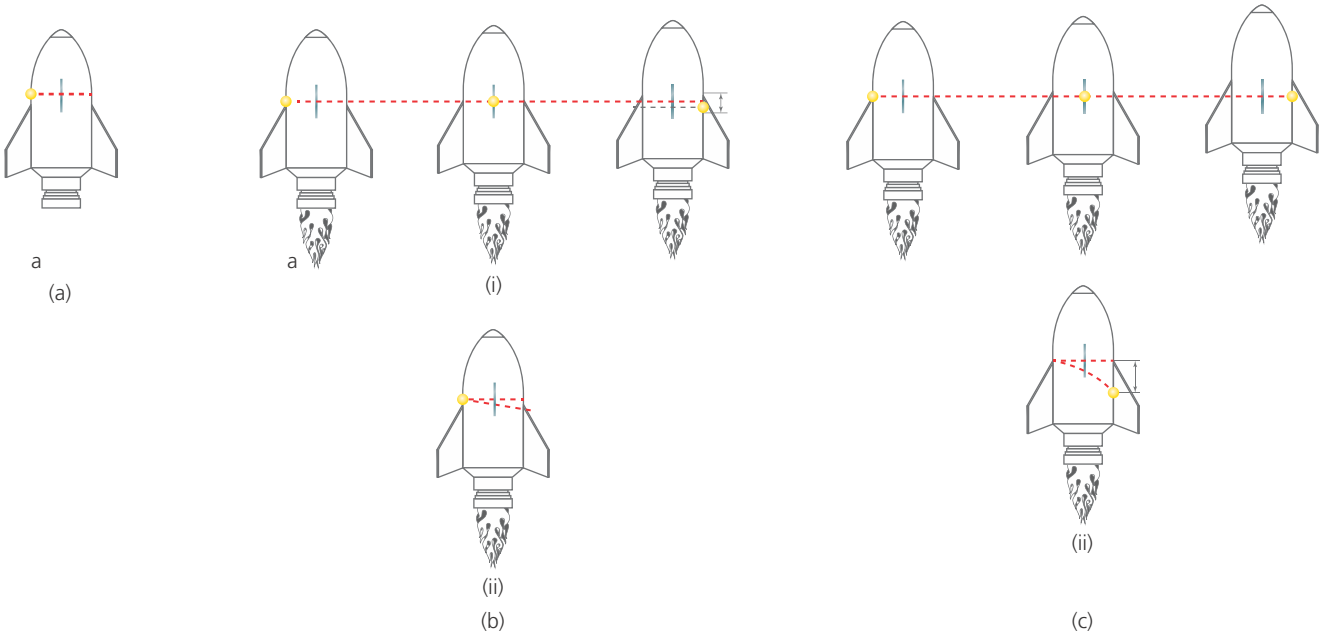


Figure 13.45 Thought experiments to show bending of light. (a) The rocket is stationary. Light travels straight horizontally (b). The rocket has a constant upward velocity. Panel (b.i) shows the straight horizontal path of light as seen from the outside. Panel (b.ii) shows the straight but sloping path of light as seen by an inside observer. (c) The rocket has a constant upward acceleration. Panel (c.ii) shows the curved path as seen by an inside observer.

- Since light could bend in an accelerating frame of reference, then from the principle of equivalence, it can be concluded that a beam of light would also follow a curved path in a gravitational field.

The bending of light

- Recall that for an object moving at constant speed, its worldline in the space-time diagram is a straight line. It is easy to see that an accelerating object will follow a curved worldline. Thus, an accelerated frame of reference will have a curved space-time, as shown in Figure 13.46.

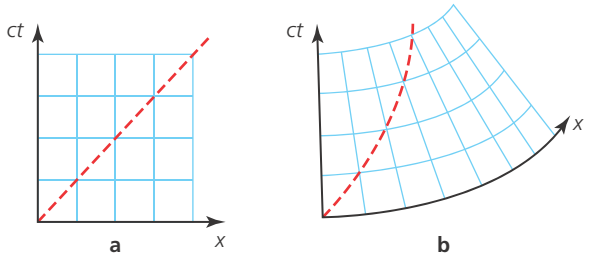


Figure 13.46 The worldline of an object in an inertial frame (a) and in an accelerated frame (b).

- According to the principle of equivalence, the curved space-time due to acceleration of a mass is also the space-time in a gravitational field.
- The bending of light can be explained by the curving of space-time. Space-time is generally flat, but the presence of mass deforms and curves space-time. The larger the mass, the greater is the curving or warping of space-time. The closer it is to the massive object, the more severe is the curvature of space-time (see Figure 13.47).
- A curved or warped space-time, with its bumps and depressions, creates what is felt as gravity. Thus, under the general theory of relativity, gravity is not a force but a manifestation of the warping of space-time due to the presence of a mass.

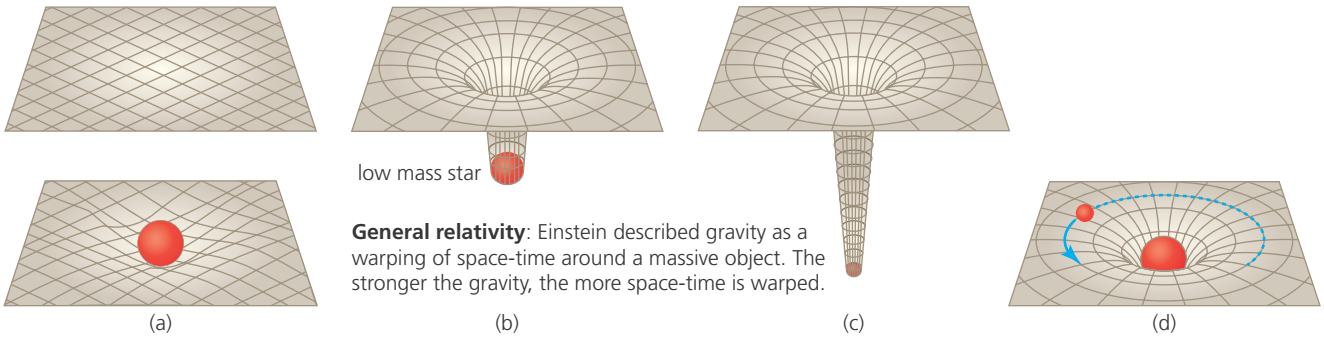


Figure 13.47 A two-dimensional representation of space-time. (a) A flat region of space-time. (b) Mass distorts space-time. (c) Distortion of space-time increases with mass. (d) A planet follows the curvature of space-time when it orbits around the Sun.

- When light travels in a curved space-time, it moves along the curve of shortest path between two points, called the 'geodesic'.
- In space-time anything, including light, moves along the curve of shortest path between two points, the geodesic.
  - As an example, the Sun curves the space-time around it. The orbit of the Earth about the Sun is the shortest path through the curved space-time around the Sun (Figure 13.47d).
  - Rather than the Earth reacting to the presence of a force field, its orbit is due to its following a curved space-time without the need to know the source of the curvature.
- Therefore, mass affects the curvature of space-time and conversely the curvature of space-time dictates how objects, including light, will move.
- With a relatively small mass, Earth's warping of space-time is not enough to produce a noticeable bending of light. However, it has been proven experimentally that the Sun and other astronomical bodies can bend light.

Equivalence principle and gravitational time dilation

Revised

- The principle of equivalence also predicted gravitational time dilation which means time runs slower in regions of stronger gravitational field.
- To illustrate how this effect arises, consider the motion of a photon emitted parallel to the acceleration of an elevator (Figure 13.48).
  - A photon of frequency f_0 is directed upward from the floor of the elevator that is accelerating upward. As the photon travels from the floor to the ceiling, the elevator gains speed. Also, the ceiling is moving away from the source. The situation is shown in Figure 13.48.
  - By the time the photon reaches the ceiling, the ceiling is moving away from the source at a higher speed than when the photon was emitted. Due to the Doppler effect, the frequency f_f of the photon that reaches the ceiling is lower than the frequency that was emitted from the source on the floor.
  - Since the speed of light is constant in any reference frame, then the period of the photon is greater at the ceiling than at the floor, e.g. period is 2.0 ns at the ceiling but only 1.0 ns at the floor. This means 2.0 ns have elapsed at the ceiling, but only 1.0 ns on the floor. Thus, time runs slow on the floor.
  - By the principle of equivalence, the same results are expected in a gravitational field. In this example, the floor represents the region of stronger gravitational field. Hence, time runs slow inside a strong gravitational field.
- Gravitational time dilation is caused by the warping of space-time. A stronger gravitational field leads to greater warping of space-time. This means that light must travel a longer path to reach two points in space-time. To compensate for the longer path, time slows down, thus keeping the speed of light constant.
- Unlike relativistic time dilation, gravitational time dilation is not symmetrical. All observers agree that time runs slower in the region of stronger gravitational field.

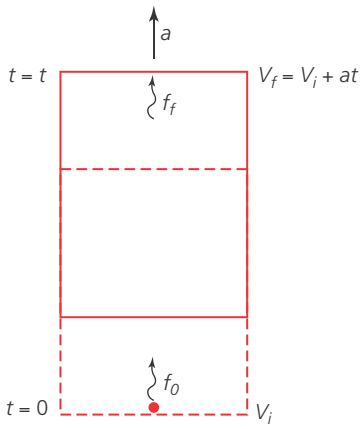


Figure 13.48

Gravitational redshift and the Pound-Rebka-Snider experiment

- **Gravitational redshift** (Figure 13.49) happens when the wavelength of a photon changes to a longer wavelength or lower frequency as it moves to a region of weaker gravitational field.
- When a photon is moving upwards, it gains potential energy. By energy conservation, it should lose kinetic energy. Since the energy of a photon is $E = hf$, then a decrease in the frequency leads to a reduction in the energy.
- Consequently, when the photon falls towards the surface, its frequency increases due to its gain in potential energy.
- The frequency change Δf due to gravitational effects can be calculated using the equation below.

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

where f is the original frequency, g is the gravitational field strength, Δh is the change in height and c is the speed of light.

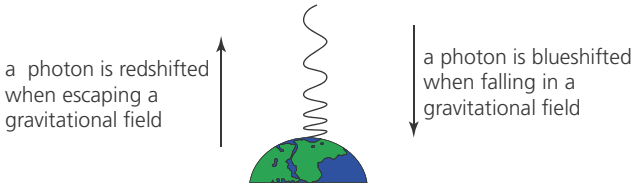


Figure 13.49

- In the early 1960s, Pound and Rebka conducted an experiment which verified gravitational redshift. Later, the experiment was repeated with even higher precision by Pound and Snider. The details of the experiment are described below.
- The experiment used the principle called the Mössbauer effect: identical nuclei emit and absorb photons of the same frequency. However, if the frequency is changed even by a small amount, then the photon can no longer be absorbed by the nuclei identical to the emitters.
- Gamma rays from radioactive nuclei (Fe-57) were directed vertically upward from the basement of the 22.5 m high Jefferson Tower at the Harvard University campus. A detector that contained the same type of radioactive nuclei was placed at the top of the tower. A diagram of the experiment is shown in Figure 13.50.
- Results showed that the detector did not absorb the photons coming from the basement of the building. This suggested that the photons had redshifted. However, calculations showed that the frequency shift was so small that physical measurement was impossible.
- To measure the frequency shift, Pound and Rebka mounted the gamma ray emitters on a loudspeaker to move them at a controlled velocity and induce a Doppler blueshift. They adjusted the oscillation of the loudspeaker to exactly cancel out the gravitational redshift and bring the gamma rays back to the correct frequency to be absorbed.
- Initially, Pound and Rebka calculated from this method the amount of gravitational redshift and this agreed with the predictions of general relativity to within 10%. Later on, Pound and Snider repeated the experiment and improved the agreement to within 1%.

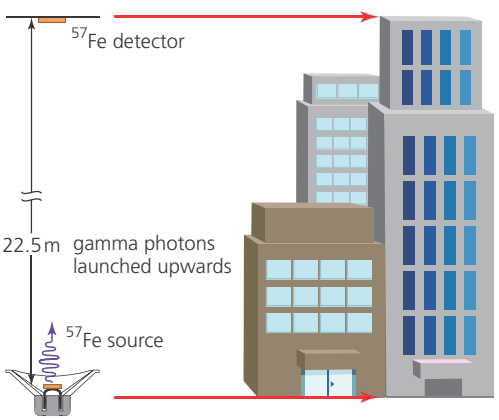


Figure 13.50 Pound-Rebka-Snider experiment.

Calculating gravitational frequency shifts

Revised

The gamma rays in the Pound-Rebka-Snider experiment had 14.4 keV of energy. The building was 23 m high. Calculate the change in the frequency when a photon falls from the top of the building.

Energy of the photon in J: $14.4 \times 10^3 \text{ eV} \times \frac{1.6 \times 10^{-19}}{1 \text{ eV}} = 2.30 \times 10^{-15} \text{ J}$

Frequency of the photon: $E = hf \rightarrow f = \frac{E}{h} = \frac{2.30 \times 10^{-15}}{6.63 \times 10^{-34}} = 3.47 \times 10^{18} \text{ Hz}$

Frequency shift: $\frac{\Delta f}{f} = \frac{g\Delta h}{c^2} \rightarrow \Delta f = \frac{fg\Delta h}{c^2} = \frac{3.47 \times 10^{18} \times 9.81 \times 23}{(3.0 \times 10^8)^2} = 8700 \text{ Hz}$

Schwarzschild black holes and event horizon

Revised

- Another prediction of the general theory of relativity is the existence of **black holes**. Some black holes are formed by the collapse of stars that are 1.4 times the mass of the Sun. Some black holes are primordial formed in the early universe and have rather small mass. Others form in the centre of giant galaxies and have very large masses.
- A black hole is a region in space-time that has extreme curvature due to the presence of a highly dense mass. A black hole is 'black' because it absorbs all the light that enters it and reflects none. The gravitational field inside the black hole is infinitely strong so that nothing, not even light, can escape from it. The extreme distortion of space-time causes photons to curve back.

- A black hole consists of two main parts: the singularity and the event horizon.
  - The centre of a black hole is called the singularity where all matter in the black hole is compressed in an infinitesimally small point of volume, resulting in an infinitely large density. At the singularity, space-time has infinite curvature and it is believed that time comes to a stop.
  - The event horizon is the surface where the escape velocity is equal to the speed of light. The event horizon marks the boundary within which nothing, not even light, can escape. Anything that crosses the event horizon disappears from the observable universe forever.
  - Contrary to popular myth, a black hole is not a cosmic vacuum cleaner, which would suck up all matter. Outside of the event horizon, the space-time is flat and the gravitational field of the black hole is the same as the gravitational field of an object of the same mass. To be pulled into the black hole, an object has to cross the event horizon.
- A Schwarzschild black hole is a static non-rotating black hole.
- The Schwarzschild radius is the radius at which a gravitationally collapsing body becomes a black hole. Once an object collapses to within its Schwarzschild radius, it will continue to collapse toward singularity. This radius is equal to the radius of the event horizon.
  - Light emitted from the Schwarzschild radius would take an infinitely long time to reach an outside observer and it would be redshifted to an infinitely long wavelength in the process.
  - Events inside the event horizon can have no causal effect on the universe outside of the horizon because no information can be sent from inside to outside of the event horizon.
- The Schwarzschild radius R_S is proportional to the mass of the black hole and can be calculated using the equation below.

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

where $G = 6.67 \times 10^{-11}\ \text{N m}^{-2}\text{kg}^{-2}$, M = mass of the black hole and c = speed of light.

Calculating the Schwarzschild radius of a black hole

Calculate the Schwarzschild radius of the Sun (mass = 1.99×10^{30} kg) if it were to become a black hole. Would the Earth and the other planets be sucked in if the Sun became a black hole?

$$R_S = \frac{2GM}{c^2} = \frac{2 \times 6.67 \times 10^{-11} \times 1.99 \times 10^{30}}{(3.0 \times 10^8)^2} = 3000\,\text{m}$$

Thus, if the Sun were to become a black hole, its Schwarzschild radius would be around 3 km. The average orbital distance of the planets from the Sun is much greater than 3 km. Since the mass of the Sun will not change during collapse (it is now infinitely dense but not more massive), the gravitational force on the Earth is still the same, as it is on the other planets. Hence, the Earth's and the other planets' orbits will remain unchanged and the planets will not be sucked in by the Sun-black hole. On the other hand, of course, the Earth's average temperature would drastically decrease.

Time dilation near a black hole

- A probe nearing the event horizon of a black hole will be seen by observers as experiencing a dramatic redshift as it gets closer, so that time appears to be going more and more slowly as it approaches the event horizon.

- Two stationary observers who are at different distances from a black hole will measure the time interval between the same two events differently.

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

where Δt = time elapsed for an observer or a clock that is distant from the black hole; Δt_0 = time elapsed for an observer or a clock that is near the black hole; r = radial distance of the 'near observer' or clock from the black hole; R_s = Schwarzschild radius.

- The above formula shows that time gets more dilated nearer the black hole. At the critical distance $r = R_s$, time is dilated infinitely. The equation has no meaning if $r < R$.
- As r gets closer to R, the event horizon, then Δt approaches infinity. This means time stops.

Applying the formula for gravitational time dilation

A black hole has a Schwarzschild radius R. A clock at a distance of $2R$ from the singularity of the black hole measures the time between two events to be 15 s. What is the time between these two events as measured by a clock that is very far from the black hole?

$$\Delta t = \frac{\Delta t_0}{\sqrt{1-\frac{R_s}{r}}} = \frac{15}{\sqrt{1-\frac{R}{2R}}} = \frac{15}{\sqrt{\frac{1}{2}}} = 21 \text{ s}$$

Applications of general relativity to the universe as a whole

Revised

The general theory of relativity was such a huge success that it became the theoretical foundation for the great cosmological revolution of the 20th century. General relativity has predicted the existence of strange wonders.

- Distant light sources behind massive objects can appear to move, or get brighter or change shape. As such, massive objects like stars and galaxies can act as gravitational lens, refocusing and magnifying light from a distant source, thereby distorting its image. This is shown in Figure 13.51.

  Gravitational lensing can also result into the formation of multiple images. Depending on the shape and geometry of the lensing mass (the nearer massive object) the resulting image could be an arc, a ring of light or a series of multiple images, as shown on Figure 13.52.

  Gravitational lensing is a powerful tool used by astronomers to study the universe. Its magnifying effect allows observers to see objects that are too far away or faint for even the largest telescopes on Earth. Many scientists believe that most of the universe is in the form of invisible dark matter and dark energy. Images of distant galaxies that appear severely distorted and magnified imply that a large amount of dark matter is located in front of them. The idea of dark matter emerged to explain discrepancies between two different ways of finding the masses of galaxies.

- Dark matter is named such because it interacts gravitationally like ordinary matter but is undetectable by emission or absorption of radiation. The idea of dark matter emerged to explain the expansion of the universe, which was initially predicted by the general theory of relativity. Hubble later verified this and even showed that the expansion is not steady but is accelerating.

Key concepts

Gravitational lensing. Gravitational lensing is an effect of the bending of light when light passes near a massive object.

Dark matter. Dark matter is a hypothetical form of matter that is believed to account for 27% of matter in the universe.

Gravitational waves. Gravitational waves are disturbances or ripples in the fabric of space-time due to violent and energetic processes in the universe like colliding black holes, exploding supernovae and coalescing neutron stars.

Calculations show that the universe does not have enough mass to cause its accelerated expansion, hence dark matter was proposed.

- Gravitational waves propagate at the speed of light, stretching and compressing space time as they travel. They were first detected on 14 September 2015 from merging black holes.

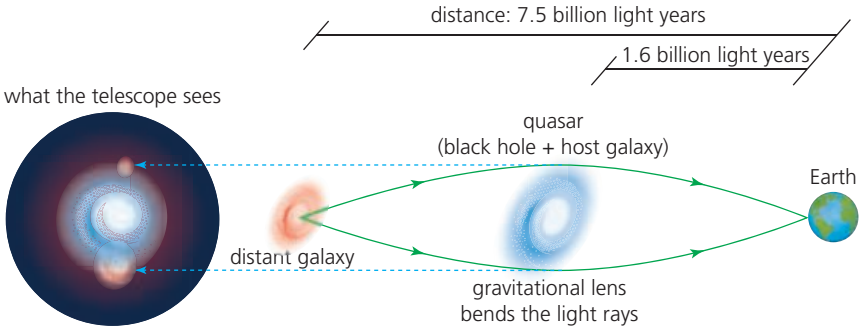


Figure 13.51 Gravitational lensing. The massive object curves space-time forming a distorted multiple image of the light source.

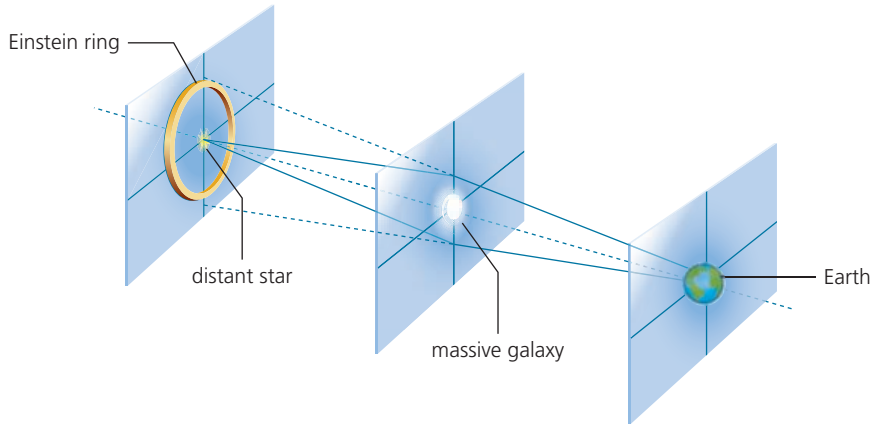


Figure 13.52 A spherical lensing mass forming a ring of light known as an Einstein ring by deforming light from a distant galaxy (or star).

Option B 14 Engineering physics

14.1 Rigid bodies and rotational dynamics

Revised

Essential idea: The basic laws of mechanics have an extension when equivalent principles are applied to rotation. Actual objects have dimensions and they require the 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.

Torque

Revised

- For *linear motion*, a (resultant) force is needed to change the motion of an object (see 2.2 Forces, p. 20). Changing *rotational motion* involves a force which is not directed towards the axis of rotation. We say that a *torque* is required.
- The concept of torque is similar to the concept of the **moment of a force**, with which students may be familiar.
- In this section on rotational dynamics, we will only discuss **rigid bodies**: objects which do not change their shapes.
- Figure 14.1 shows an example of three equal forces that might be used with a spanner. Force F_1 has no turning effect because its **line of action** is through the *axis of rotation*. Force F_2 has the greatest possible turning effect because its line of action is the furthest from the axis of rotation.

Key concept

A (resultant) **torque**, Γ, is needed to change the *rotational* motion of an object.

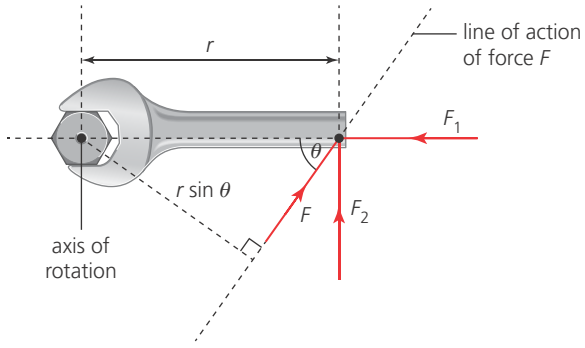


Figure 14.1

Calculating torque for single force and couples

- In general, the torque provided by a force F, which has a line of action which 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 Nm.

Common mistake

Torque, $Fr\sin\theta$ (unit Nm), should not be confused with work, $Fs\cos\theta$ (Chapter 2) which also has the unit Nm, but which is more commonly called the joule, J.

Key concept

The torque provided by a single force can be determined from $\Gamma = Fr\sin\theta$.

A **couple** is the name given to a pair of parallel, equal magnitude forces which have different lines of action and act in opposite directions, tending to cause rotation (see Figure 14.2).

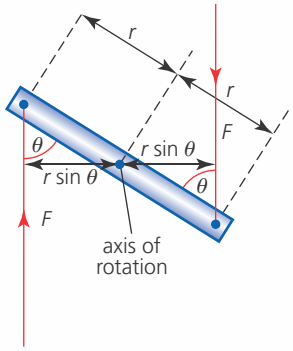


Figure 14.2

- 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.
- (Students may be familiar with the **principle of moments**: an object will remain in equilibrium if the sum of the clockwise moments acting on it equals the sum of the anticlockwise moments.)
- The torque provided by a couple is the sum of the two individual torques (which are often equal to each other).
- Using two hands to turn a wheel is an everyday example of a couple. Examples from elsewhere in this course include the forces on a rotating coil, or a bar magnet, in a magnetic field (see Figure 14.3 which shows that as the magnet rotates, the torque reduces and becomes zero when the magnet is aligned with the field.)

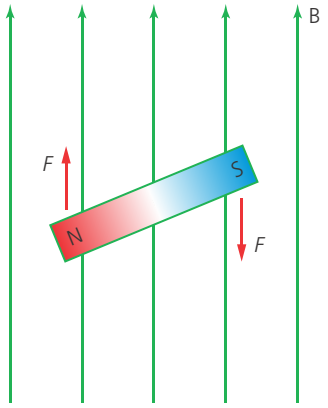


Figure 14.3

Moment of inertia

Revised

- In linear motion, resistance to a change of motion is known as **inertia**. Larger masses have greater inertias.
- For rotational motion, resistance to a change of 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 example is a point mass, m, which is a distance r from its axis of rotation, as shown in Figure 14.4. Its moment of inertia can be determined from $I = mr^2$. The unit of moment of inertia is kg m^2.
- Note that whenever an equation is needed in an examination for the moment of inertia of a particular shape, it will be provided in the question.
- As an example of a non-point mass, consider a thin rod of length L and mass m as shown in Figure 14.5. Because there are different axes of rotation, the same rod would be easiest to rotate in (a) and hardest in (c). (The moment of inertia for Figure 14.5b is $\frac{mL^2}{12}$ and for Figure 14.5c, it is larger, $\frac{mL^2}{3}$).
- If necessary, the overall moment of inertia of more complicated shapes can be determined by adding the moments of inertia of the parts. Question 7 shows an example.

Key concept

The moments of inertia of all masses can (in principle) be found from summing the moments of inertia of all their points: $I = \Sigma mr^2$.

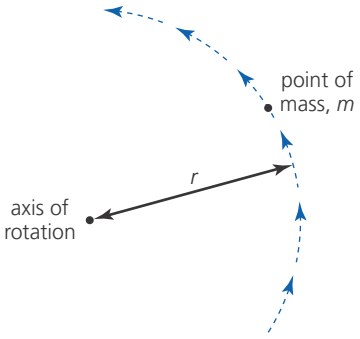


Figure 14.4

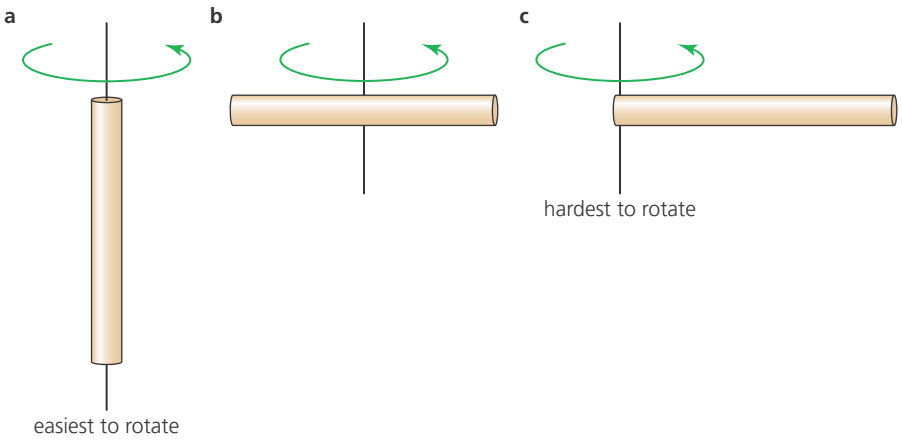


Figure 14.5

QUESTIONS TO CHECK UNDERSTANDING

1 Consider Figure 14.1. If the point of application of a force was 16 cm from the axis of rotation
   a what torque would be provided if $F = 72$ N and $\theta = 58°$?
   b What is the maximum possible torque from a force of 120 N at the same point?
2 Give two examples of couples (other than those mentioned above).
3 Calculate the torque provided by the couple shown in Figure 14.2 if the two forces were both 100 N, r was 8.3 cm and θ was 68°.
4 A 1.2 kg mass hangs vertically on the end of a 3.3 m string. The mass is free to swing from side to side.
   a What is its moment of inertia?
   b What assumption did you make?
5 Consider the rotation of the thin rod shown in Figure 14.5a. Apart from its mass and length, suggest what other information would be needed in order to determine its moment of inertia.
6 The moment of inertia of a solid sphere about an axis through its centre is $\left(\frac{2}{5}\right)mr^2$.
   a Determine the moment of inertia of a solid sphere of radius 5.0 cm and mass 0.38 kg.
   b Suggest what material such a sphere could be made of.
   c What would be the moment of inertia of a solid sphere of the same material which had twice the radius?
7 A 'dumbbell' shape is produced if two solid spheres are added to the rotating thin rod shown in Figure 14.5b. See Figure 14.6. Each sphere has a moment of inertia of mr^2 (considered to be point masses), where r is the distance from the centre of the sphere to the axis of rotation. Determine the overall moment of inertia of this shape if the length of the rod is 40 cm and it has a mass of 64 g, and each sphere has a mass of 0.35 kg and radius of 3.1 cm.

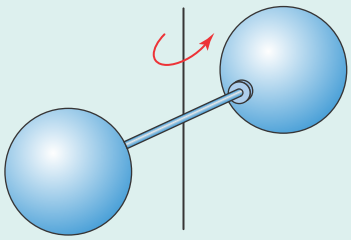


Figure 14.6

Expert tip

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.

Rotational and translational equilibrium

Revised

- From Newton's first law of motion in Chapter 2, Section 2.2: An object is in **translational equilibrium** if it is stationary or moving with constant linear velocity. In other words, an object in translational equilibrium has zero acceleration. Translational equilibrium occurs when there is no resultant force acting on an object.
- Rotational equilibrium can be defined in a similar way.
- An object may be in translational equilibrium and not rotational equilibrium, or in rotational equilibrium but not translational equilibrium, or it may be in both types of equilibrium (or neither).
- If we consider an object to be a **point particle**, we can easily represent the action of a resultant force in a simple drawing, with only one outcome possible, as shown in Figure 14.7.

Key concept

An object is in **rotational equilibrium** if it is rotating with a constant angular velocity (including being at rest). This occurs when there is no resultant torque acting on it.

- However, as soon as we try to represent an object more realistically, such as in Figure 14.8, we realise that a force can result in changes to both translational and rotational motion, unless its line of action is through the centre of mass. Putting spin on a ball struck with a tennis racket is a good example of this.
- In this course, numerical examples will only involve objects revolving around fixed axes of rotation.

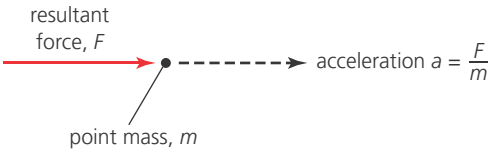
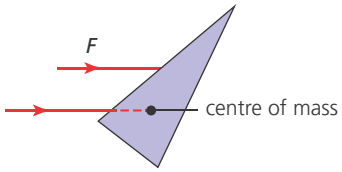


Figure 14.7



The result of a force F depends on its line of action

Figure 14.8

Reminders from Chapter 6, Section 6.1, Circular motion

- *Angular velocity*, $\omega = \frac{\Delta\theta}{\Delta t}$ (where θ is *angular displacement*, usually measured in radians). Angular velocity has the unit rad s^{-1}.
- For an object moving in rotational equilibrium, with a constant linear speed v in a circular path of radius r, such that its period is T and its frequency is f:
  - $\omega = \frac{\text{one rotation}}{\text{one period}} = \frac{2\pi}{T}$, or $\omega = 2\pi f$ (because $T = \frac{1}{f}$).
  - Since $v = \frac{2\pi r}{T}$, $v = \omega r$.

Solving problems in which objects are in both rotational and translational equilibrium

QUESTIONS TO CHECK UNDERSTANDING

8 Give an example of:
- **a** an object in translational equilibrium but not in rotational equilibrium, and
- **b** an object in rotational equilibrium and translational equilibrium.

9 An object spinning at a constant rate moves through an angle of 80° in 0.73 s.
- **a** What is this angle in radians?
- **b** Calculate its angular velocity.

10 A cylindrical mass is rotating about its central axis with a constant angular velocity of $540\ \text{rad s}^{-1}$.
- **a** What is its period?
- **b** What is its frequency?
- **c** If the cylinder has a radius of 2.2 cm, what is the speed of
  - **i** a point on its circumference,
  - **ii** a point midway between the circumference and the centre?

Angular acceleration

Revised

- A resultant torque on an object will produce an angular acceleration, α, so that it will not be in rotational equilibrium.
- *Angular acceleration* can be compared to linear acceleration $a = \frac{\Delta v}{\Delta t} = \frac{(v-u)}{\Delta t}$
- Since $v = \omega r$, angular acceleration and the linear acceleration, a, of a point which is a distance r from the axis of rotation are linked by the equation: $\alpha = \frac{a}{r}$.

Key concept

Angular acceleration, $\alpha = \frac{\Delta\omega}{\Delta t}$

or $\alpha = \frac{(\omega_f - \omega_i)}{\Delta t}$ Unit: rads^{-2}

Equations of rotational motion for uniform angular acceleration

Revised

It should be clear that the mathematics 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 uniform angular accelerations:

$$\theta = (\omega_f + \omega_i)\frac{t}{2}$$

$$\omega_f = \omega_i + \alpha t$$

$$\theta = \omega_i t + \frac{1}{2}\alpha t^2$$

$$\omega_f^2 = \omega_i^2 + 2\alpha\theta$$

QUESTIONS TO CHECK UNDERSTANDING

11 **a** What is the final angular velocity of a rotating object which accelerates from rest for 3.2 s with an angular acceleration of π rad s^{-2}?

b The object then slows down at a constant rate to an angular velocity of 2.4 rad s^{-1} in 7.3 s. Calculate its acceleration.

12 A point on the circumference of a wheel accelerates from 2.8 m s^{-1} to 12.3 m s^{-1} in 8.4 s. Calculate:

a the linear acceleration of that point

b the angular acceleration of the wheel if its radius is 18 cm.

13 Write down the equations of linear motion which are analogous to the equations for rotational motion highlighted above.

14 A spinning disc accelerates from 4.0 rad s^{-1} to 9.3 rad s^{-1} in 3.9 s. Through what angle does it move in this time in:

a radians

b degrees?

15 A wheel decelerates at 0.87 rad s^{-2} for 4.5 s. If its final angular velocity was 2.3 rad s^{-1}, what was the initial angular velocity?

16 An object is rotating at 200 rpm (revolutions per minute), when a torque applied to it produces an angular acceleration of 12.0 rad s^{-2}. Through what angle does it rotate in the next ten seconds?

17 A spinning fairground ride is rotating at 1.2 rad s^{-1}. What angular acceleration will stop the movement in five complete rotations?

18 A wheel is rotating at an angular velocity of 8π rad s^{-1}. If it is then accelerated at 5.0 rad s^{-2}, what will its final angular velocity be after it has moved through 20 complete rotations?

Sketching and interpreting graphs of rotational motion

Graphs of rotational motion can also be drawn and interpreted by analogy with linear motion. See the details shown on Figure 14.9.

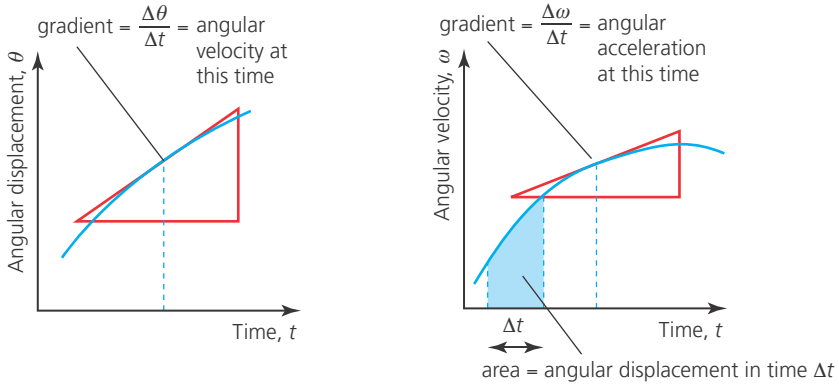


Figure 14.9

Newton's second law applied to angular motion

Revised

From Chapter 2, Section 2.2, we know that $F = ma$ (Newton's second law for linear motion). A similar equation applies to angular motion.

Key concept

Torque = moment of inertia × angular acceleration $\Gamma = I\alpha$.

Solving problems involving moment of inertia, torque and angular acceleration

Worked example

Figure 14.10 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.

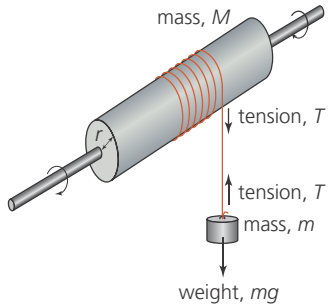


Figure 14.10

Torque acting on cylinder, $\Gamma = Tr = I\alpha = \frac{1}{2}Mr^2 \times \frac{a}{r}$

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 $a = \frac{mg}{m + \frac{1}{2}M}$

QUESTIONS TO CHECK UNDERSTANDING

19 Figure 14.11 shows how the angular velocity of a rotating object changed over a time of 20 s.

a Describe the motion represented by the graph.

b Determine the acceleration of the object in the last 8 s.

c Estimate the total angle through which the object rotated in 20 s.

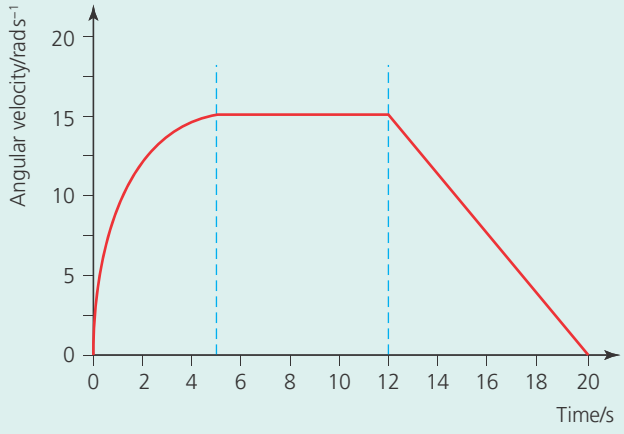


Figure 14.11

20 a What is the moment of inertia of an object which is accelerated by 20 rad s^{-2} by a resultant torque of 25 Nm?

b What torque is required to bring the same object to rest from an angular velocity of 60 rad s^{-1} in 8.0 s?

21 A hollow sphere has a moment of inertia of $\frac{2}{3}mr^2$. What angular acceleration will be produced when a tangential force of 10 N acts on a basketball of radius 12.4 cm and mass 625 g?

Conservation of angular momentum

Revised

- From Chapter 2, Section 2.4, we know that for linear motion, momentum (mass × linear velocity) is conserved in every isolated system (a system with no external forces acting on it).
- By analogy, angular momentum, L (moment of inertia × angular velocity) is conserved in rotational motion.
- This means that if the moment of inertia of an isolated system is changed, then its angular velocity must also change. Figure 14.12 shows an example: the ice skater can increase her moment of inertia by extending her arms; this will result in an angular deceleration as her angular velocity is reduced. Similarly, if a child jumps onto a rotating unpowered carousel in a park playground, the greater mass will increase the moment of inertia and slow the ride down.

Key concept

Angular momentum, $L = I\omega$.
Unit: kg m^2 s^{-1}

The total angular momentum of a system is constant provided that no resultant external torque is acting on it.

Figure 14.12

Changing angular momentum

- Also from Chapter 2, Section 2.4, we know that impulse = force × time = change of linear momentum. This too, has a rotational equivalent:
  - Torque × time = change of angular momentum: $\Gamma \times \Delta t = \Delta L$.
- Figure 14.13 shows an example in which the torque changes. The change of angular momentum can then be determined from the area under the graph. The unit N s for linear impulse has a rotational equivalent of N m s.

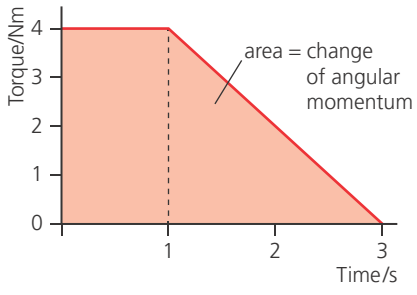


Figure 14.13

QUESTIONS TO CHECK UNDERSTANDING

22 A 20 g rubber ball on the end of a length of string is rotated at a constant speed in an (almost) horizontal circle of radius 76 cm. If the ball completes ten rotations in 8.47 s, determine:

a its angular velocity

b its angular momentum.

23 A wooden disc has a moment of inertia of 1.38×10^{-3} kg m^2. It was rotating horizontally with negligible friction and a constant angular velocity of 11.3 rad s^{-1} when a 100 g mass was carefully dropped onto its surface, 8.5 cm from the centre.

a What is the moment of inertia of the extra mass?

b Determine the new angular velocity of the disc.

24 Determine the change of angular momentum represented by the graph in Figure 14.13.

Rotational kinetic energy

Revised

- Rotational kinetic energy, E_{Krot}, can be determined from an equation similar in form to that for linear kinetic energy $\left(E_K = \frac{1}{2}mv^2\right)$.

Key concept

Rotational kinetic energy can be calculated from the equation $E_{Krot} = \frac{1}{2} I \omega^2$.

Objects which have rotational and translational energy

- Many objects have both rotational and translational kinetic energy. The wheels on a moving vehicle and a ball rolling down a hill are obvious examples. On the microscopic scale, gas molecules rotate as they move from place to place.

Key concept

Objects may have both linear and rotational kinetic energy:

$E_{Ktotal} = \frac{1}{2}mv^2 + \frac{1}{2} I \omega^2$

Solving problems involving rolling without slipping

- An object that can **roll**, like a wheel or a ball, needs friction in order to start to roll. Without friction it will **slip**.
- When an object, like a wheel, is rolling (without slipping) the point on the wheel which is in contact with a stationary surface must have zero resultant velocity at that instant (see Figure 14.14).
- If a wheel or ball is rolling with an angular velocity of ω, we know that the linear speed of a point on the circumference $v = \omega r$ and this is also the translational speed of the centre of the wheel (or ball).
- Figure 14.15 shows a ball rolling down a hill. The gravitational potential energy at the top of the slope will be transferred to the two forms of kinetic energy, as above.

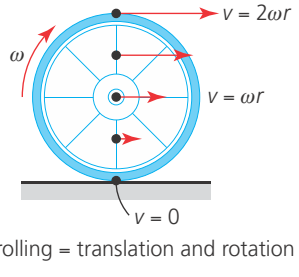


Figure 14.14

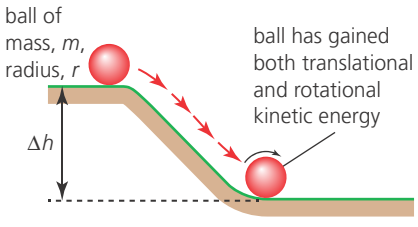


Figure 14.15

- As an example, a solid sphere (ball) of mass m and radius r has a moment of inertia, $I = \left(\frac{2}{5}\right)mr^2$, so that $E_{Ktotal} = \frac{1}{2}mv^2 + \frac{1}{2} I \omega^2$ reduces to $E_{Ktotal} = \frac{7}{10} m \omega^2 r^2$. The angular velocity at the bottom of the slope can be determined by equating this to $mg\Delta h$.

QUESTIONS TO CHECK UNDERSTANDING

25 Calculate the total kinetic energy of a tennis ball of mass 60 g that is spinning at a rate of four revolutions per second as it moves through the air with a translational speed of 10 m s^{-1}. (Moment of inertia = 5.0×10^{-5} kg m^2)

26 A car is moving at 50 km h^{-1} (13.9 m s^{-1}). If the outer radius of the wheel and tyre is 22 cm, what is

a the linear speed of a point on the circumference

b the angular velocity of the wheel?

27 a Show that the total kinetic energy of a sphere of mass m and radius r rolling along horizontal ground (without slipping) with an angular velocity of ω is $\left(\frac{7}{10}\right) m\omega^2 r^2$, given. $I = \left(\frac{2}{5}\right) mr^2$

b If the sphere had a radius of 14.20 cm and rolled from rest down a slope of vertical height 24 cm, determine its

i angular velocity

ii linear velocity.

Comparing linear and rotational concepts and equations

Revised

Table 14.1 Summary of the concepts and equations for linear and rotational mechanics

Linear mechanics	Rotational mechanics
displacement, s	angular displacement, θ
initial velocity, $u = \frac{\Delta s}{\Delta t}$	initial angular velocity, $\omega_i = \frac{\Delta \theta}{\Delta t}$
final velocity, v	final angular velocity, ω_f
acceleration, $a = \frac{\Delta v}{\Delta t}$	angular acceleration, $\alpha = \frac{\Delta \omega}{\Delta t}$
$v = u + at$	angular velocity, $\omega = \frac{2\pi}{T} = 2\pi f$
$s = \left(\frac{u+v}{2}\right) t$	$\omega_f = \omega_i + \alpha t$
$s = ut + \frac{1}{2} at^2$	$\theta = \left(\frac{\omega_i + \omega_f}{2}\right) t$
$v^2 = u^2 + 2as$	$\theta = \omega_i t + \frac{1}{2} \alpha t^2$
	$\omega_f^2 = \omega_i^2 + 2\alpha\theta$
force, F	torque, $\Gamma = Fr \sin\theta$
mass, m	moment of inertia, $I = \sum mr^2$
$F = ma$	$\Gamma = I\alpha$
linear momentum, $p = mv$	angular momentum, $L = I\omega$
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, $E_{K_{rot}} = \frac{1}{2} I\omega^2$

Expert tip

Further analogies are possible, although they are not needed in this course. For example, in linear mechanics, work done is Fs, and in rotational motion, work done is $\Gamma\theta$.

Solving problems using rotational quantities analogous to linear quantities

- The analogies seen in Table 14.1 have been used in questions throughout this section.

NATURE OF SCIENCE

Modelling

The use of a point particle model for any object acted on by forces greatly simplifies many situations, so that analysis and making predictions becomes much easier. However, the single point model cannot be used when forces do not act through the centres of mass of objects, or when there is a fixed axis about which they can rotate. In such cases, we can extend the model to consider the object as a collection of different points.

14.2 Thermodynamics

Revised

Essential idea: The first law of thermodynamics relates the change in internal energy of a system to the energy transferred and the work done. The entropy of the universe tends to a maximum.

- The study of thermodynamics in this chapter concentrates on the transfer of thermal energy to expanding gases in order to do useful work. An understanding of thermodynamics requires knowledge of the kinetic theory of gases from Chapter 3.
- Figure 14.16 shows a common visualization and simplification of this type of process: in this example the gas is contained in a cylinder by a sliding piston which is able to move (without friction) when there is a difference in gas pressure between its inner and outer surfaces.

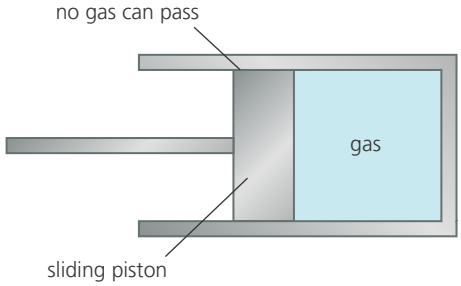


Figure 14.16

Heat engines

Revised

- When thermal energy is transferred into the gas in Figure 14.16, the molecules move faster and the piston is pushed outwards, so that the pressure inside and outside is kept the same.
- An expanding gas does work as it pushes back the surrounding air.
- This important principle is used to transfer thermal energy to useful mechanical work in vehicle engines and power stations throughout the world. The practical details may vary, but all such devices are known as **heat engines**.
- The earliest useful forms of heat engines used steam to do the work (see Figure 14.17) and this is still true in the latest power stations.
- Figure 14.18 shows the energy flow principle of heat engines: thermal energy is made to flow from hot to cold and some of that energy is used to do useful work. We will see later that it is not possible for all of the thermal energy to be transferred to useful work.
- The gas is often known as the 'system', everything else is called the 'surroundings'.
- The piston shown in Figure 14.16 obviously cannot keep moving outwards in an infinite cylinder, so in order to be practicable, the gas in a heat engine

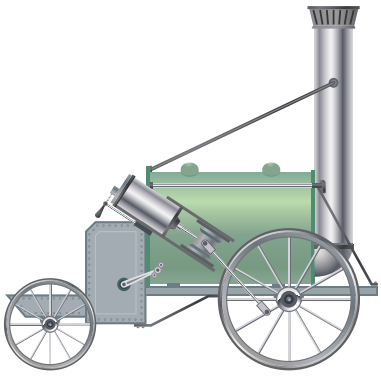


Figure 14.17

Key concept

The expansion of a heated gas is widely used to do useful work in various types of engines. Such 'heat engines', working in rapidly repeating cycles, are used in most vehicles and for the generation of electricity.

must be compressed after its expansion, so that the process can be repeated. This means that practicable heat engines must work in repeating **cycles** (more details below).

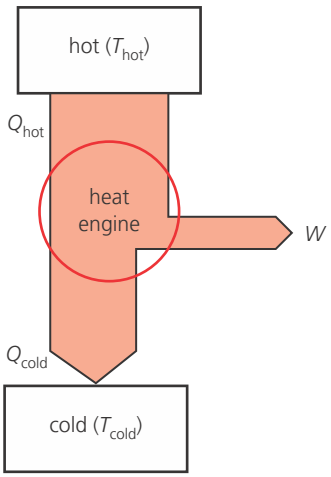


Figure 14.18

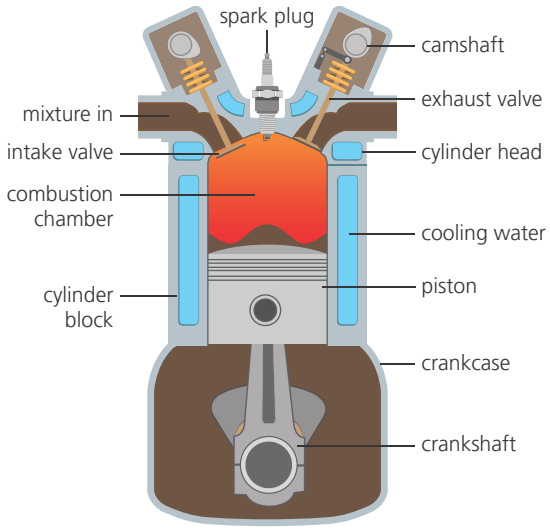


Figure 14.19 Internal combustion engine.

Expert tip

Figure 14.19 shows a piston in the cylinder of an internal combustion engine, typically used in cars. Such an engine has a four stroke cycle, but it is presented here for interest only. No knowledge of engine details is needed for the IB Physics course.

The physical properties of gases (mostly revision from Chapter 3)

Revised

- The physical state of a fixed amount of a gas can be fully described by three variables: pressure, p, volume, V and absolute temperature, T.
- The equation of state for an ideal gas can be used to predict what happens when a gas changes state: $pV = nRT$.
- Real gases behave like ideal gases under most circumstances. Although it should be noted that most real gases contain molecules which have rotational kinetic energy as well as translational energy. (This means that the equation for internal energy, shown in the box on the right, needs to be adapted, but this is not in the IB Physics course.)
- The state of a gas is commonly represented on a pressure–volume graph, with different curves representing different temperatures, as shown in Figure 14.20. These are known as pV diagrams. The lines connecting points at the same temperature are called 'isotherms'. Each one is representing the behaviour of a gas consistent with Boyle's law.

Key concept

The internal energy of an ideal gas is just the translational kinetic energy of its particles. It is given the symbol U and can be calculated from $U = \frac{3}{2}nRT$. (n is the amount of gas in moles.)

Key concept

***pV* diagrams** are used to represent gas behaviour at different temperatures.

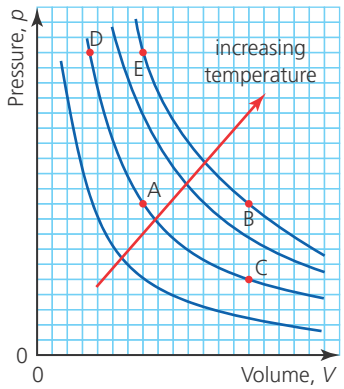


Figure 14.20

- 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 can only happen at a higher temperature. Moving from A to C also represents a doubling of volume, but at a constant temperature, so that the pressure must be lower. Moving from A to point D can be achieved by halving the volume at the same temperature so that the pressure doubles. Finally, moving from A to E represents doubling the pressure by increasing the temperature in the same volume.

QUESTIONS TO CHECK UNDERSTANDING

28 An ideal gas exerts a pressure of 7.9×10^4 Pa on the walls of its container which has a volume of 25 cm^3. If the temperature was –8.0°C, how much gas was in the container?

29 a Sketch a *pV* diagram with one isothermal line for a temperature *T*.

b A gas starts at one point on the isothermal and its volume expands by 150% at constant pressure. It is then cooled at constant volume until its temperature returns to *T*. It then returns at constant temperature to its original volume. Sketch this process (cycle) on your *pV* diagram.

30 Give four examples of different kinds of heat engine.

31 Outline the differences between ideal gases and real gases.

32 a Calculate the internal energy of one mole of an ideal gas at 273 K.

b In what form is the internal energy of an ideal gas?

c How much energy is needed to raise 4.0 moles of an ideal monatomic gas from 48°C to 73°C?

Work done when a gas changes volume

Revised

- When a gas expands, work is done *by the gas* against the pressure exerted by the surroundings. In this course, the work done is considered to be *positive*. When a gas is compressed the work is done *on the gas* and is considered to be *negative*.
- Figure 14.21 shows the expansion of this gas at constant pressure. We know that in general, work done = force × distance moved in the direction of the force (Chapter 2), so that for a gas, work done = $(pA)\Delta s$ (because $F = pA$). Because $\Delta V = A\Delta s$ we have $W = p\Delta V$. The same equation is valid for any shaped container.
- In the example shown in Figure 14.22, the pressure of the gas is *not* constant; it is expanding as its pressure decreases and the shaded area represents the work done by the gas on its surroundings.

Key concept

The (maximum) work done, *W*, when a gas changes volume can be determined from the area under a *pV* diagram. If the pressure remains constant, we can simply use $W = p\Delta V$.

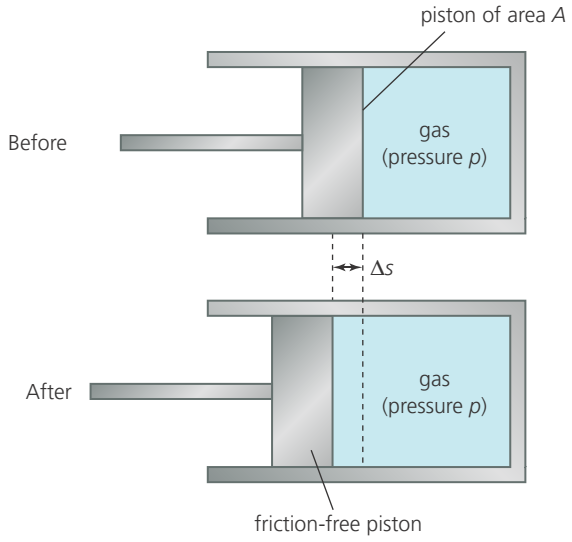


Figure 14.21

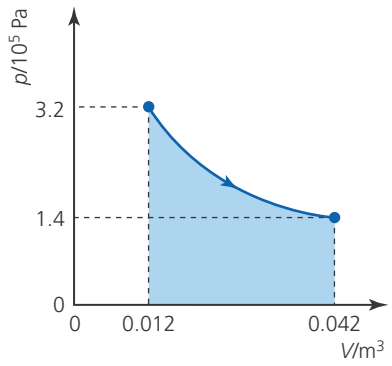


Figure 14.22

QUESTIONS TO CHECK UNDERSTANDING

33 Some gas is contained in a cylinder of cross-sectional area 8.72 cm^2.

- **a** When thermal energy is transferred into the gas, it expands at a constant pressure of 1.45×10^5 Pa, pushing the piston 2.3 cm along the cylinder. Determine the work done by the gas in this process.
- **b** Explain why the piston moves back along the cylinder when the gas is cooled.

34 A gas at a pressure of 1.0×10^5 Pa has a volume of 16 cm^3. It is compressed at constant temperature to a pressure of 4.0×10^5 Pa and volume of 4 cm^3.

- **a** Sketch a *pV* diagram for this process.
- **b** Use your sketch to estimate the work done while the gas was being compressed.

The first law of thermodynamics

Revised

- The meanings of **thermal energy** and **internal energy** need to be well understood (Chapter 3).
- In general, in any thermodynamic process, we may consider that if a quantity of *thermal energy*, $+Q$, is supplied to a gas, the gas may get hotter so that the *internal energy* of the gas may rise by an amount $+\Delta U$ and/or the gas may expand to do *work*, $+W$.

Describing the first law of thermodynamics as a statement of conservation of energy

- Using the principle of conservation of energy (Chapter 2), it should be clear that the thermal energy supplied to a gas must equal the increase in its internal energy plus the work done by the gas if it expands. This is known as the first law of thermodynamics.

Key concept

The **first law of thermodynamics**:

$Q = \Delta U + W$

The signs used with these quantities must be considered carefully.

Explaining sign convention used when stating the first law of thermodynamics

- Of course, it is possible that thermal energy is *removed* from a gas and its internal energy and/or volume could *decrease*. When using the first law of thermodynamics each of these processes must be given a negative sign.

Solving problems involving the first law of thermodynamics

QUESTIONS TO CHECK UNDERSTANDING

35 340 J of thermal energy were transferred into a gas while it expanded doing 450 J of work.

- **a** What was the change in internal energy of the gas?
- **b** Did the gas get hotter or colder?

36 620 J of work were done when a gas was compressed. At the same time, its temperature went up so that its internal energy rose by 280 J.

- **a** Determine how much thermal energy was transferred.
- **b** Was the thermal energy transferred into or out of the gas?

Isovolumetric, isobaric, isothermal and adiabatic processes

Revised

- A gas may change state in many different ways but, in practice, a change of state usually approximates to one of the four idealized processes represented in Figure 14.23.

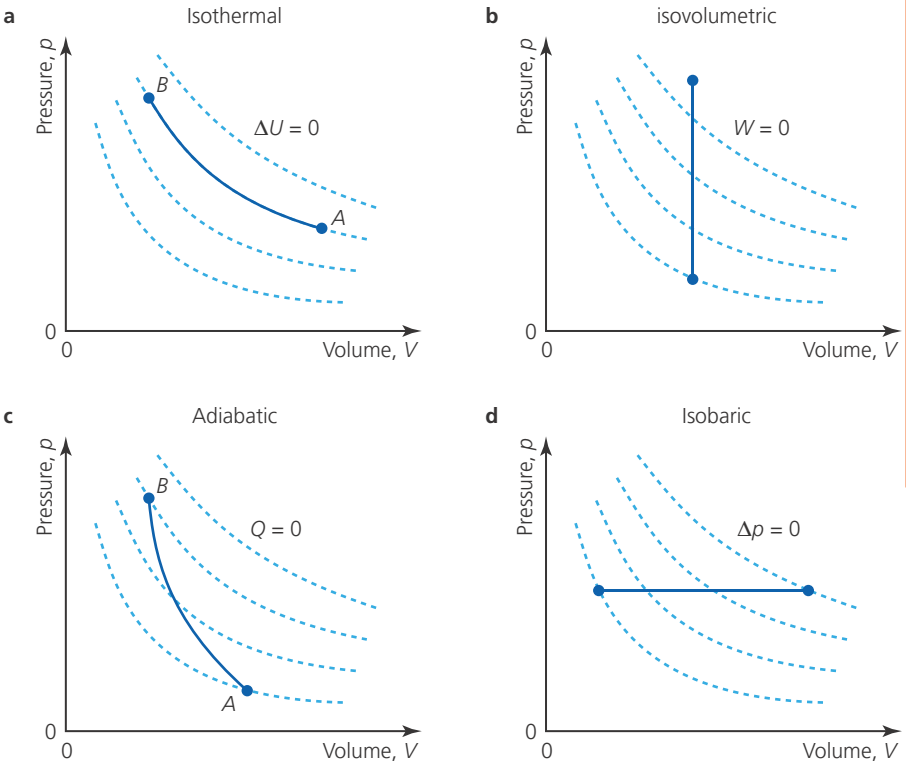


Figure 14.23

Key concepts

Isovolumetric processes occur at constant volume, so that there is no work done in expansion or compression, $W = 0$.

Isobaric processes occur at constant pressure, $\Delta p = 0$.

Isothermal processes occur at constant temperature, so that there is no change to the internal energy of the gas, $\Delta U = 0$.

Adiabatic processes occur when no thermal energy is transferred to or from the gas, $Q = 0$.

- From an *isovolumetric change* $Q = \Delta U + W$ becomes $Q = \Delta U$. All of any thermal energy transferred changes only the internal energy of the gas. A simple example would be heating or cooling a gas in a well-insulated container of fixed dimensions.
- *Isobaric changes* occur when gases are allowed to expand or contract freely when their temperature changes, keeping their pressure the same as the surrounding pressure.

  $Q = \Delta U + W$.
- For an *isothermal change* $Q = \Delta U + W$ becomes $Q = W$. All of any thermal energy transfer is used as work to expand or compress the gas.

  From Chapter 3: pV = constant (Boyle's law).
- In an isothermal expansion (B → A in Figure 14.23a), all of 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 *adiabatic change*, $Q = \Delta U + W$ becomes $-\Delta U = W$ or $\Delta U = -W$. All of the energy required for an expansion (B → A in Figure 14.23c) is taken from the internal energy of the gas, which must cool ($-\Delta U = W$). Conversely, the work done on a gas which is compressed adiabatically (A → B), is transferred completely to the internal energy of the gas, which must get hotter ($\Delta U = -W$).

- On the microscopic scale, in an adiabatic process, kinetic energy is transferred between the moving piston and the molecules which collide with it.
- Many changes to gas volume can approximate to adiabatic processes if they occur quickly enough that there is not enough time for significant thermal energy transfer into or out of the gas (especially if the container is well insulated).

Solving problems for adiabatic changes for monatomic gases using the equation $pV^{\frac{5}{3}} = \text{constant}$

- Comparing graph a to graph c (Figure 14.23), it should be noted that, for equal changes of volume of the same gas at the same starting pressure, there is a greater change in pressure for an adiabatic process than an isothermal process. This is because of the changes to temperature and average molecular speed that occur in adiabatic processes.
- We know that $pV\,(= pV^1) = \text{constant}$ for isothermal processes. For adiabatic processes, $pV^{\gamma} = \text{constant}$, where for monatomic gases, $\gamma = \frac{5}{3}$.
- For example, consider a monatomic gas in a container of volume $V_1 = 0.40\,\text{m}^3$ at a pressure of $p_1 = 2.0 \times 10^5\,\text{Pa}$. If the gas expands *isothermally* to a volume $V_2 = 0.60\,\text{m}^3$, the new pressure $p_2 = \frac{p_1 V_1}{V_2} = 1.3 \times 10^5\,\text{Pa}$. However, if the same gas had expanded *adiabatically* to the same volume, the final pressure would be $p_2 = p_1 \left(\frac{V_1}{V_2}\right)^{\frac{5}{3}} = 1.0 \times 10^5\,\text{Pa}$.

Key concepts

For *isothermal* changes of ideal gases, $\boldsymbol{pV = \text{constant}}$.

For *adiabatic* changes of ideal monatomic gases (like helium, neon and argon) $pV^{\frac{5}{3}} = \text{constant}$.

Expert tip

Most gases are molecular and therefore not monatomic. For these gases, the constant γ has different values because their molecules have rotational as well as translational energies. Details of this are not needed in the IB Physics course.

QUESTIONS TO CHECK UNDERSTANDING

37 The volume of a gas was halved and, as a result, its pressure doubled in an isothermal process.
- **a** What happened to the temperature of the gas?
- **b** Apply the first law of thermodynamics to this process.
- **c** The gas was then returned adiabatically to its original volume.
  - **i** Represent these two changes on a *pV* diagram.
  - **ii** Compare the starting and final temperatures of the adiabatic change and account for the difference, if any.

38 The volume of a gas expanded from $5.9\,\text{cm}^3$ to $8.1\,\text{cm}^3$ while its pressure remained constant at $1.0 \times 10^5\,\text{Pa}$.
- **a** What name do we give to this kind of process?
- **b** Calculate the work done by the gas.
- **c** 1.0 J of thermal energy was transferred into the gas during the process.
  - **i** Determine the change of internal energy.
  - **ii** Did the gas get hotter or colder?

39 Explain how the pressure of a gas can be made to increase during an isovolumetric process.

40 Helium gas of volume $0.174\,\text{m}^3$ and pressure $3.3 \times 10^5\,\text{Pa}$ expands adiabatically until its pressure is $1.7 \times 10^5\,\text{Pa}$.
- **a** What is the final volume of the gas?
- **b** Under what conditions can we assume that this expansion was approximately adiabatic?
- **c** What happened to the temperature of the helium?

41 In diesel engines, the air and fuel mixture is ignited by the process of rapidly compressing the gas. Use your understanding of adiabatic changes to explain this process.

Cyclic processes and *pV* diagrams

Revised

- As already mentioned, heat engines involve repeated cycles which can be represented as closed figures on *pV* diagrams.

Sketching and interpreting cyclic changes

- A discussion of Figure 14.24 shows how such cycles can be interpreted, but it has been simplified to a rectangular shape and is not intended as a practical example. Between A and B the pressure is increasing at constant volume as thermal energy is supplied to the gas (an isovolumetric process). Between B and C the gas is doing useful work as it expands in an isobaric process, but since the pressure remains constant, we know that thermal energy is still being supplied, increasing molecular speeds. C to D is another isovolumetric process, so no work is done, and since the pressure is reduced, we know that thermal energy has been extracted from the gas. If work is done on the gas and thermal energy is extracted in an isobaric process from D to A, the system is then returned to its original state and the cycle can then be repeated.
- The area under the line BC represents the work done by the expanding gas. The area under DA represents the work done on the gas during compression. The (shaded) area enclosed by the cycle represents the net useful work done during the cycle. In order to extract useful work from a thermal energy input the process must occur in a clockwise direction around the cycle.

Key concept

The repeating cycle of a heat engine designed to do useful work from an input of thermal energy can be represented by a clockwise closed path on a *pV* diagram.

The enclosed area represents the useful energy transferred in each cycle.

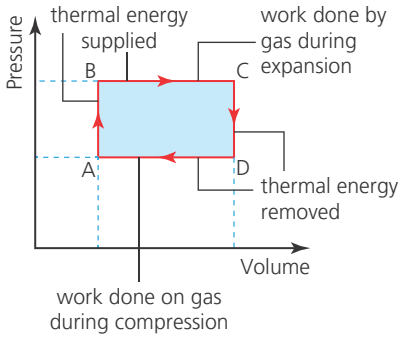


Figure 14.24

Expert tip

Air conditioners and refrigerators are also heat engines, but they work in the opposite sense: they do mechanical work to transfer energy from a colder place to a hotter place. They are known as **heat pumps**. Heat pumps can also be represented by cycles on *pV* diagrams, but in an anticlockwise direction.

Carnot cycle

- The Carnot cycle (see Figure 14.25) 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.)
- If an examination question involves a thermodynamic cycle other than the Carnot cycle, full details will be provided.

Key concept

Theoretically, the most efficient thermodynamic cycle is known as the **Carnot cycle**. Thermal energy is transferred into the gas during isothermal expansion and removed during isothermal compression.

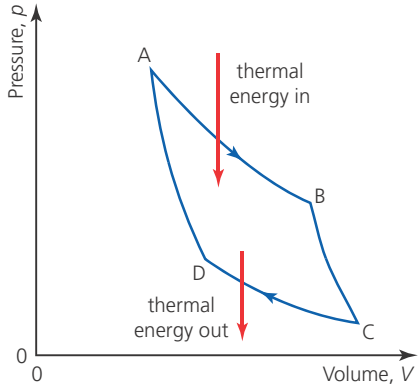


Figure 14.25

QUESTIONS TO CHECK UNDERSTANDING

42 During a Carnot cycle 4.0×10^{-3} mol of a gas at a pressure of 9.3×10^{5} Pa expands isothermally in volume from 18 cm^3 to 43 cm^3.

- **a i** What was the temperature of the gas?
- **ii** Calculate its final pressure.
- **b** The gas then undergoes an adiabatic expansion to a volume of 75 cm^3 while its pressure falls to 1.2×10^{5} Pa. This is followed by an isothermal compression to a volume of 34 cm^3, before it returns adiabatically to its original state. Sketch the complete cycle on a *pV* diagram.
- **c** Use your diagram to estimate the useful work done during each cycle.

43 Figure 14.18 represents the energy flow of a heat engine doing useful work. Draw a similar diagram to represent a heat pump.

Thermal efficiency

Revised

- We have seen in Chapter 2 that, in general, the efficiency of any process,

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

However, the efficiency of heat engines, and in particular a Carnot cycle, can be related to temperatures.

- If there is no temperature difference, there is no flow of thermal energy and the efficiency is zero. If the outlet temperature could be absolute zero, the efficiency would be 1.
- The outlet temperature may be assumed to be similar to the temperature of the surroundings, for example 300 K, and this has important implications that limit the efficiencies of heat engines.
- For a fixed outlet temperature, increasing the inlet temperature will improve efficiency, but such increases are limited by the temperature-dependent properties of the materials used in the engine.

Key concept

Thermodynamic efficiency of a Carnot cycle depends on the temperatures involved:

$$\eta_{\text{Carnot}} = 1 - \frac{T_{\text{cold}}}{T_{\text{hot}}}$$

where T_{hot} is the inlet temperature of the engine (hot **reservoir**) and T_{cold} is the outlet temperature (cold reservoir).

Solving problems involving thermal efficiency

QUESTIONS TO CHECK UNDERSTANDING

44 The total power of thermal energy transfer in a heat engine is 2.5 kW. If it operates at an efficiency of 30%, how much useful mechanical work is done every minute?

45 What is the maximum theoretical efficiency of a heat engine operating between temperatures of:

a 320 and 650 K?

b 280 and 610 K (the same difference)?

46 Outline why it is not possible for fossil-fuelled power stations to ever be more than about 50% efficient, but hydro-electric power stations can be over 90% efficient.

Entropy

Revised

- Left to themselves, ordered things naturally become disordered. External influences are necessary to create order from disorder.
- 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.
- Consider Figure 14.26 which shows possible arrangements of gas molecules in a container at different times. Because it is so unlikely for molecules moving randomly, we simply cannot believe that C could occur *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 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 naturally becomes more disordered.

Key concept

The random and uncontrollable behaviour of molecules will always result in increasing molecular disorder.

Entropy is a numerical measure of the disorder of a system.

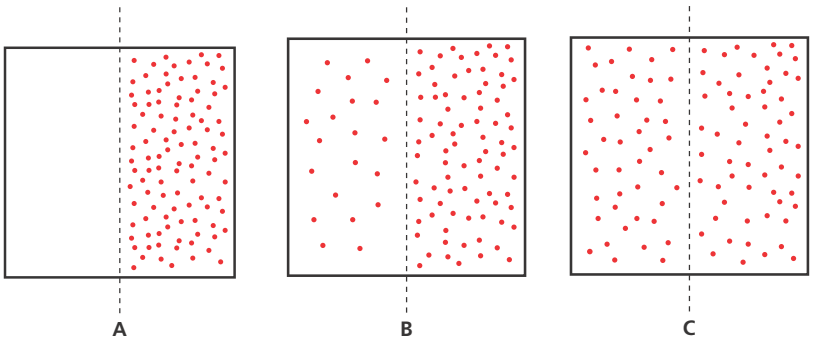


Figure 14.26

- Because of this process, useful energy always becomes degraded as it gets disordered and dissipated into the surroundings. (It cannot be recovered later to do any useful work.)
- This inevitable process of molecular **order** changing to **disorder** is so fundamental to an understanding of physics that it requires a concept to describe it: *entropy*.
- There is no requirement to calculate absolute entropy values in the IB Physics course.

The second law of thermodynamics

Revised

- The second law of thermodynamics summarizes ideas about disorder and energy dissipation.

Describing examples of processes in terms of entropy changes

- When a cup of hot coffee left on a table cools down its entropy decreases as thermal energy is dissipated from the coffee into the surroundings, but the entropy of the surroundings increases more than the entropy of the coffee decreased.
- When a cold drink warms up thermal energy is absorbed from the surroundings and the entropy of the surroundings decreases less than the entropy of the cold drink increases.
- It is possible to artificially reduce the entropy of part of a system, for example by freezing water, but the energy dissipated from that part of a system will always increase the entropy of the surroundings even more.

Key concept

Second law of thermodynamics: *all* processes increase the entropy of an *isolated system* (and the universe as a whole).

Key concept

All energy transfers involve entropy changes. Typically, some parts of a system will decrease in entropy while other parts increase in entropy. However, the *total* entropy always increases.

Describing the second law of thermodynamics in Clausius form, Kelvin form and as a consequence of entropy

- Apart from the entropy interpretation of the second law of thermodynamics (above), the underlying principle can also be stated in other, more practical ways.

Key concepts

Alternative statements of the second law:

1. When extracting energy from a heat reservoir, it is impossible to convert it all into work (**Kelvin form**);
2. Thermal energy cannot spontaneously transfer from a region of lower temperature to a region of higher temperature (**Clausius form**).

Solving problems involving entropy changes

- For example, suppose that 1000 J of thermal energy was transferred from an object at 350 K to an object at 300 K. Consider the idealized situation in which the objects are insulated from their surroundings and the exchange of energy had no significant effect on their temperatures. The change of entropy of the hotter object was $-\frac{1000}{350} = -2.86\,\text{J K}^{-1}$ and the change of entropy of the cooler object was $+\frac{1000}{300} = +3.33\,\text{J K}^{-1}$. The overall change of entropy is $+0.48\,\text{J K}^{-1}$.

  There has been an increase of entropy, as predicted by the second law of thermodynamics, which also confirms that thermal energy is transferred from the hotter to the cooler object.

Key concept

The *change* in entropy of a system, ΔS, when thermal energy ΔQ is added or removed at a constant temperature T (K) (in a reversible change), can be calculated from $\Delta S = \frac{\Delta Q}{T}$. The units of entropy (change) are J K^{-1}.

QUESTIONS TO CHECK UNDERSTANDING

47 Discuss the energy and entropy changes when a ball rolling on a horizontal surface is brought to rest by friction.

48 Use the equation $\Delta S = \frac{\Delta Q}{T}$ to discuss the Clausius form of the second law of thermodynamics.

49 a Calculate the energy that has to be transferred to melt 100 g of ice at 0°C. The latent heat of fusion of water is $3.3 \times 10^5\,\text{J kg}^{-1}$.

b Determine the entropy change of the ice in this process.

c If the ice was melted by placing it in a large mass of warm water at 30°C, estimate the overall entropy change of the ice and water system.

NATURE OF SCIENCE

Variety of perspectives

The three varying (but equivalent) statements of the same law, the second law of thermodynamics, illustrate that basic physics principles expressed concisely may need to be interpreted in different ways for different contexts. Newton's second law of motion is another example: sometimes we use $F = ma$, but on other occasions an interpretation in terms of momentum is preferred.

14.3 Fluids and fluid dynamics (Additional higher level)

Revised

Essential idea: 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.

Density and pressure

Revised

- The term **fluid** is used to describe substances which can flow: liquids or gases.
- If a gas which is *free to expand* is heated, its volume will increase, so that at a higher temperature it has a lower density (same mass). If it is *not* able to expand when it is heated, its density will remain unchanged, but the pressure will increase.
- Most liquids expand slightly when they are heated, so that there will be a small decrease in density.
- A fluid cannot be considered as a point, so any force acting will usually be spread over an area, so that we usually refer to the *pressures* in and on fluids, rather than forces.
- The **atmospheric pressure** at sea level is 1.0×10^5 Pa, which is equivalent to a force of 10 N on each square centimetre.

Key concepts (revision)

Density = $\frac{\text{mass}}{\text{volume}}$; $\rho = \frac{m}{V}$
(unit: kg m^{-3})

Pressure = $\frac{\text{force}}{\text{area}}$; $p = \frac{F}{A}$
(unit: N m^{-2} or Pa)

Solving problems involving pressure and density

QUESTIONS TO CHECK UNDERSTANDING

50 If the density of air is 1.18 kg m^{-3}, calculate the mass of air in a room of dimensions $3.2 \times 2.3 \times 4.7$ m.

51 If the air in a car tyre exerts a pressure of 2.5×10^5 Pa, what is the force acting on an area of 150 cm^2 (the approximate area of a tyre in contact with the ground)?

Pressure in fluids

Revised

- A fluid can experience a pressure because of an externally applied pressure (including from the atmosphere) and/or pressure due to its own weight.
- Because of the random motions of its molecules, the pressure at a point in a fluid acts equally in all directions, including upwards.
- The pressure in a stationary fluid due (only) to its own weight can be determined by using $p = \frac{F}{A}$. This is an example of **hydrostatic pressure**.
- Consider Figure 14.27, which represents a column of liquid of depth d and density ρ_f.
- Pressure = $\frac{\text{weight of fluid}}{\text{area } A}$. $p = \frac{A\rho_f gd}{A}$ or $p = \rho_f gd$.

Key concept

The pressure due to a depth d of a fluid can be calculated from $p = \rho_f gd$. This pressure acts equally in all directions.

Common mistake

Although this calculation has been for fluid (in a container) with vertical sides, the same equation applies to a fluid column of any shape: *fluid pressure is independent of the shape of any container in which the fluid is placed*. This is demonstrated in Figure 14.28, in which the pressures at the bottom of each tube are the same (or else the liquid would move until it was).

Figure 14.28

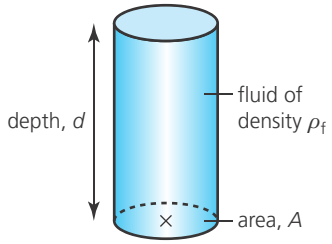


Figure 14.27

- If a point is under two (or more) fluids, the total pressure is found from the sum of the individual pressures.
- The depth of a column of liquid can be used to measure pressure. Figure 14.29 shows a U-tube containing a liquid. The tube at B is connected to the surrounding atmosphere and the tube at A is connected to the gas whose pressure is to be determined.
- The pressure at level C must be the same in both tubes, which means that the pressure of the gas, p, is equal to atmospheric pressure + $\rho_f g \Delta h$. A more detailed analysis would have to take the pressure due to the depth of gases in the tubes into account, but these are often assumed to be negligible.
- A U-tube used to measure pressure in this way is often called a *manometer*.

Key concept

Most commonly, the pressure under a liquid exposed to the atmosphere, $P = P_0 + \rho_f g d$, where P_0 is atmospheric pressure.

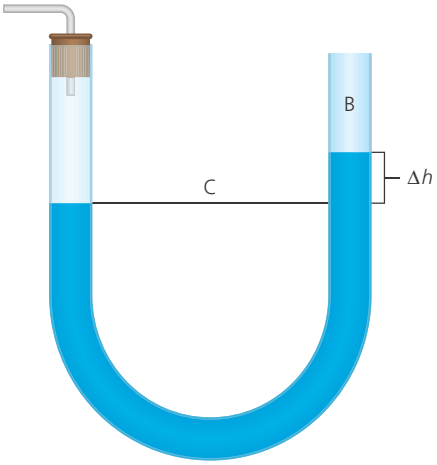


Figure 14.29

QUESTIONS TO CHECK UNDERSTANDING

52 a The troposphere is the name given to the lower part of the Earth's atmosphere which contains the densest gas. If its average height is 12 km, estimate its average density if it creates a pressure on the Earth's surface of 1.0×10^5 Pa.

b Estimate the air pressure at the top of Mount Everest (height = 8848 m).

53 a What is the height of a vertical column of water which will create a pressure equal to atmospheric pressure? Assume density of pure water = 1.00×10^3 kg m^{-3}.

b How far under the surface of sea water would a diver need to go before the pressure on him was three times the pressure at the surface? Assume sea water density = 1.03×10^3 kg m^{-3}.

54 The pressure of gas in a container was measured by the difference in levels in a manometer as in Figure 14.29. If the liquid was mercury (density 1.35×10^4 kg m^{-3}) and the height difference of the levels was 8.70 cm, what was the pressure (Pa) of the gas?

Buoyancy and Archimedes' principle

Revised

- When an object is placed completely in a fluid (*immersed*) or on its surface, some of the fluid must be *displaced*. The object experiences the same pressures as the fluid before it was displaced.
- Since pressure increases with depth, the bottom of the object experiences a greater pressure than the top. This results in a net upwards force on the object.

- This ability of a fluid to provide a vertical upwards force on any object placed in or on it is called **buoyancy**. This buoyancy force is sometimes called **upthrust**.
- The magnitude of the buoyancy force, B, can be determined by considering Figure 14.30.
- B = extra pressure × area = $\rho_f g d \times A$, or since volume $V = dA$, $B = \rho_f V_f g$. A regular shaped object has been used to obtain this equation, but it is valid for all shapes. $\rho_f V_f g$ is the weight of the fluid displaced.
- The object shown in Figure 14.31 will rise if $B > mg$ and it will fall (sink) if $B < mg$ (as shown). If a solid object is 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.
- An object which floats, like the boat shown in Figure 14.32, will move lower into the water until the weight of the water displaced equals the weight of the boat.

Key concept

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. $B = \rho_f V_f g$. This is known as **Archimedes' principle**.

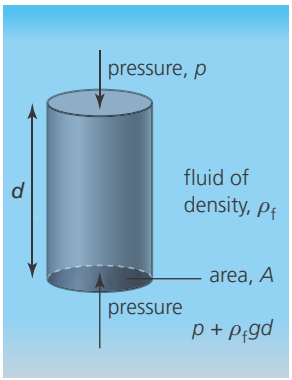


Figure 14.30

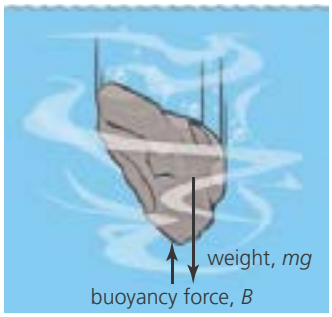


Figure 14.31

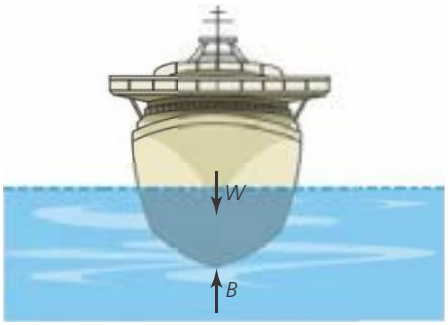


Figure 14.32

Determining buoyancy forces using Archimedes' principle

QUESTIONS TO CHECK UNDERSTANDING

55 a Calculate the buoyancy force (upthrust) on a gas-filled balloon of volume 5000 cm^3 when it is in:

i air

ii water.

b Explain why the same balloon will fall when placed in air, but rise when placed in water.

56 Determine the resultant force on a stone of mass 420 g and density 3.7×10^3 kg m^{-3} when it is immersed in water.

57 Explain why you will tend to move upwards in a swimming pool if you take a deep breath of air.

58 A solid wooden sphere of radius 6.0 cm^3 is placed on the surface of water.

a Explain why you might expect it to float.

b Use $V = \left(\frac{4}{3}\right)\pi r^3$ to determine the volume of the sphere.

c If the density of the wood is 870 kg m^{-3}, determine the volume of water displaced.

d Draw a sketch of the sphere floating on the water.

Expert tip

When a part of a fluid is heated, its density can decrease and this will result in an upthrust from the surrounding fluid. This causes the warmer part of the fluid to rise and produces a *convection current*, as discussed in Chapter 8, Section 8.2.

Pascal's principle

Revised

- Because any liquid is incompressible and its molecular motions are random, when extra pressure is applied at one point in a liquid, the same extra pressure will occur at all other points in the same liquid.
- **Hydraulic machinery** uses liquids in pipes (usually flexible) to transfer pressure. Since additional pressure $\left(p = \frac{F}{A}\right)$ is constant, the magnitude of forces can be increased (or decreased) by choosing different areas.
- This is illustrated in Figure 14.33. A force F_1 acting on a piston of area A_1 on the cylinder on the left hand side creates an increased pressure which is transferred to the right hand side creating an increased force F_2 acting on a piston of area A_2.

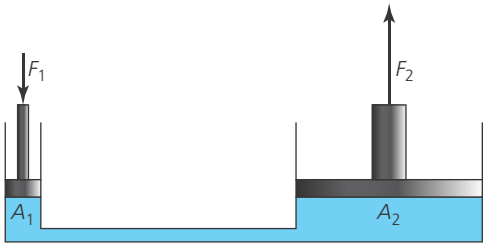


Figure 14.33

- Since the pressure is equal everywhere (Pascal's principle), $p = \frac{F_1}{A_1} = \frac{F_2}{A_2}$. So that, in principle, any force can be changed into any other larger (or smaller) force by suitable choice of the areas.
- Multiplying a force with 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% efficient. This is similar in principle to other 'force multiplying' machines like simple levers, pulleys, jacks or ramps.
- If the machine is 100% efficient (which, of course, is idealized): F_1 × distance moved by F_1 = F_2 × distance moved by F_2.

Key concept

Pascal's principle: A change of pressure exerted anywhere in an enclosed static liquid will be transferred equally to all other parts of the liquid.

Expert tip

Most cars use hydraulic brakes. Pressing the brake pedal with a relatively small force on a small area creates a pressure which is transferred by oil in flexible pipes to the braking systems on the four wheels. The force is increased by using larger areas for the pressure to act on the brakes themselves.

Solving problems involving Pascal's principle

QUESTIONS TO CHECK UNDERSTANDING

59 A hydraulic lift was used in a garage to raise a car of mass 1450 kg a height of 1.50 m. Figure 14.33 shows the principle.

a If the areas of the two cylinders were 540 cm² and 3.8 cm², what force F_1 is needed to raise the car?

b If the process was 80% efficient, what distance would the force need to move to raise the car?

Fluid dynamics

Revised

- So far in this section we have only considered stationary fluids (*hydrostatics*), we will now extend the discussion to fluids in motion: **fluid dynamics**.
- As with objects in motion, the concept of *equilibrium* is an important starting point, and we also need to consider what we mean by an *ideal fluid*.

Hydrostatic equilibrium

- **Hydrostatics** is the study of fluids at rest.
- We have previously met the concepts of translational equilibrium and rotational equilibrium. Hydrostatic equilibrium is a similar concept.

Key concept

A fluid is in **hydrostatic equilibrium** if it is either at rest or, if any parts of it are moving, they are moving with constant velocity.

The ideal fluid

- Of course, different fluids have different physical properties, but in order to study fluids in motion it is necessary to first make some assumptions about the behaviour of an 'ideal' fluid. No real fluid will have all the properties of an ideal fluid, but the concept is an essential starting point in fluid dynamics.
- We can visualize the flow of a fluid as a flow of layers capable of sliding over each other, as shown in Figure 14.34. This called *laminar flow*.
- **Incompressible** means that its volume cannot be reduced, so that density is constant.

Key concept

An **ideal fluid** is incompressible and its flow is non-viscous (viscosity is explained later). In an ideal fluid, 'layers' flow steadily without any resistance to relative motion between them. This is described as **laminar flow**.

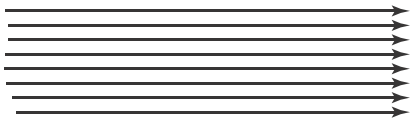


Figure 14.34

Streamlines

Revised

- The layers of flow in a liquid (see Figures 14.34 and 14.36) are generally called *streamlines*. Streamlines cannot cross each other.
- Streamlines are usually drawn with arrows showing the direction of flow from left to right. A tangent to streamlines shows the velocity of flow at that point.
- The way in which fluids flow around objects can be investigated by observing streamlines in *wind tunnels* (or liquid flow equivalents). Figure 14.35 shows smoke used to mark the streamlines around a tennis ball. The ball was probably kept stationary while high speed air was forced to move past it.

Key concept

Streamlines are lines which show the paths that (massless) objects would follow if they were placed in the flow of a fluid.

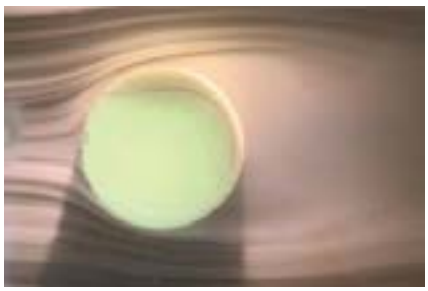


Figure 14.35

Expert tip

The flow of air around structures like buildings and bridges can be investigated using scale models in **wind tunnels** or water tanks. However, it is important to realize that the effect of scaling on fluid velocities, densities and viscosities also need to be considered.

The continuity equation

- If streamlines get closer together in a narrow section of an enclosed system, as shown in Figure 14.36, it means that the fluid must have a greater velocity, v, since the volume passing any point every second must be constant.
- This behaviour is represented by the *continuity equation*.

Key concept

Continuity equation:
Av = **constant**, where A is the cross-sectional area at any point in the enclosed system. The constant is called the **volume flow rate**.

Solving problems involving the continuity equation

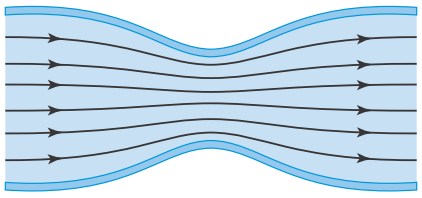


Figure 14.36

QUESTIONS TO CHECK UNDERSTANDING

60 **a** Oil is flowing through a pipeline with a speed of 0.23 m s^{-1}. What is the volume flow rate at a point where the internal diameter of the pipe is 17 cm?

b If the pipe narrowed to a diameter of 15 cm, what would the speed of the oil become?

61 At one location, a river is 24 m wide, with an average depth of 8.7 m.

a If the water is flowing at an average speed of 58 cm s^{-1}, what is the volume flow rate?

b Further downstream, the same river has widened to 31 m and the average depth is 7.2 m. What is the new average speed of the water?

c What assumption did you make in answering (b)?

The Bernoulli equation

Revised

- The *Bernoulli equation* mathematically describes the *steady flow* of an *ideal fluid* of density ρ in any enclosed system. It is based on the conservation of energy within the system (but IB Physics students are not expected to explain its origin).
- In general, we would expect that the speed of flow of a fluid in an enclosed system would change if (1) some kind of pump was providing a pressure difference; (2) the pipe was going up or down to a different level; (3) the cross-sectional area of the system was changing (as discussed above).
- Figure 14.37 shows an example: a fluid moving to a greater height and a narrower tube.

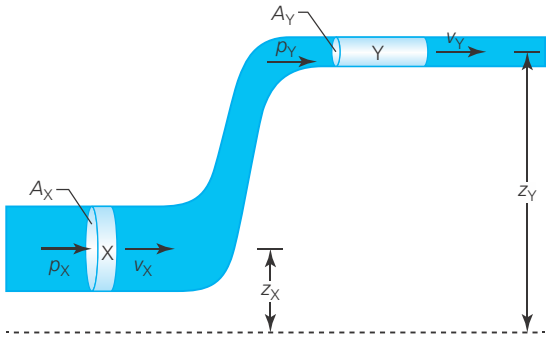


Figure 14.37

- Comparing the energy at any two points, such as X and Y, shows that:

$$\frac{1}{2}\rho v_x^2 + \rho g z_x + p_x = \frac{1}{2}\rho v_y^2 + \rho g z_y + p_y$$

The Bernoulli equation contains three terms: a term related to pressure, a term related to the kinetic energy of the fluid and a term related to the gravitational potential energy of the fluid.

- The Bernoulli equation is more usually written as $\frac{1}{2}\rho v^2 + \rho g z + p = \text{constant}$.

Key concept

The **Bernoulli equation** describes the flow of an ideal fluid through a confined system. It is usually written in the form

$$\frac{1}{2}\rho v^2 + \rho g z + p = \text{constant}.$$

This equation may be applied to the flow of a fluid out of a container, Pitot tubes and Venturi tubes.

Applications of the Bernoulli equation

- **Flow out of a container.** Consider Figure 14.38, which shows water flowing out of three holes each at a different depth below the surface. The air pressure at the three holes can be assumed to be effectively the same as the air pressure on the top surface of the water.
- Assuming that the water has zero kinetic energy at the top, the Bernoulli equation reduces to $\frac{1}{2}\cancel{\rho} v^2 = \cancel{\rho} g z$, where v is the horizontal speed of the emerging water and z is the depth of a hole beneath the surface.
- This is similar to an equation for a falling mass (Chapter 2). The different parabolic trajectories of the water in the figure clearly show how the velocity of the outlet increases with depth below the surface.

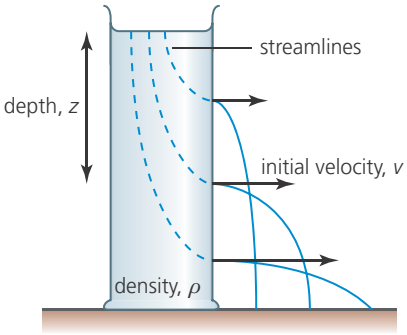


Figure 14.38

- **Pitot tubes.** Pitot tubes are used for measuring the speed of fluid flow, or the speed of an object through a fluid. Figure 14.39 shows the principle. The speed of the fluid at X is v_X but we can assume that the speed is reduced to zero after passing into the tube at Y ($v_Y = 0$).
- Using the Bernoulli equation $\left(\frac{1}{2}\rho v_x^2 + \rho g z_x + p_x = \frac{1}{2}\rho v_y^2 + \rho g z_y + p_y\right)$, ρ and z are effectively constant in this application, so that the equation reduces to $\frac{1}{2}\rho v_x^2 = +p_y - p_x = \Delta p$. In this example, Δp can be determined from the height Δh, which then enables v_X to be determined.

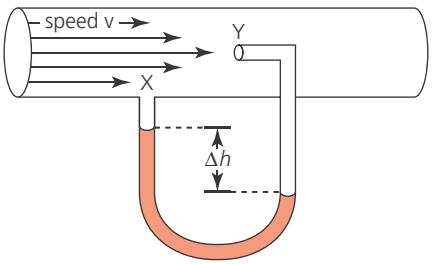


Figure 14.39

- There are many different designs of Pitot tubes. Figure 14.40 shows two Pitot tubes mounted on the side of an aircraft which are used to determine the speed of the plane relative to the air (not the land).
- Figure 14.41 shows the internal arrangement, with a pressure transducer used to measure the pressure difference between the two separate inlets.
- **Venturi tubes** are designed to produce a decrease in pressure by narrowing a section of a tube. See Figure 14.42. As before, for fluids in which any height difference is insignificant, the Bernoulli equation reduces to $\frac{1}{2}\rho v^2 + p = \text{constant}$. This equation shows us that if an (ideal) fluid flows faster, there must be a decrease in the pressure it exerts.
- In Figure 14.42, the lower height of the water in the central tube indicates that the pressure is lower where the fluid is flowing faster (because the main tube is narrower).

Figure 14.40

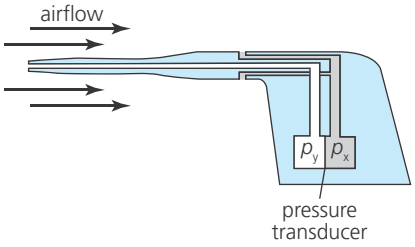


Figure 14.41

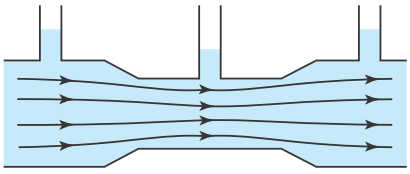


Figure 14.42

Solving problems involving the Bernoulli equation

QUESTIONS TO CHECK UNDERSTANDING

62 A small hole of area 3.0 mm^2 was drilled in the bottom of a water container.
- **a** Calculate the maximum possible speed with which the water could emerge from the hole if the water depth was 54 cm.
- **b** What is the volume flow rate out of the hole?
- **c** What radius of hole would produce a maximum volume flow rate of 20 cm^3 s^{-1}?
- **d** Suggest why, in practice, you would expect the flow rates to be lower than calculated.

63 Show that the units of $\frac{1}{2}\rho v^2$ and p are the same.

64 Consider Figure 14.39. A steady flow of gas of density 2.4 kg m^{-3} along the pipe produces a difference of water levels in the water manometer of 4.6 cm. What is the speed of the gas?

65 **a** Explain how the design of the Pitot tube shown in Figure 14.41 enables the air speed of a plane to be determined.
- **b** A plane is flying at an altitude where the air pressure is 0.47×10^5 Pa and the density is 0.69 kg m^{-3}. If the pressure on a surface perpendicular to air flow is 0.58×10^5 Pa, determine the speed of the plane through the air.

66 Consider Figure 14.42.
- **a** If a liquid of density 1050 kg m^{-3} is flowing through the pipe with a volume flow rate of 24 cm^3 s^{-1}, what is the speed of the liquid in the central section if the cross-sectional area is 0.37 cm^2?
- **b** Determine the drop in pressure that occurs in the central section if the other sections of the pipe both have cross-sectional areas of 0.81 cm^2.
- **c** What is the height difference between the liquid levels in the tubes?

67 Consider Figure 14.37.
- **a** A liquid of density 840 kg m^{-3} is passing through area A_X (6.7 cm^2) with a speed of 72 cm s^{-1}. Determine the speed of the fluid at Y if A_Y = 2.9 cm^2.
- **b** If the difference in level of the pipes is 20 cm, what pressure difference is needed to maintain the flow?

Expert tip

Remember that the Bernoulli equation only applies accurately to the steady flow of an ideal liquid (with no viscosity and no turbulence). Conclusions drawn from calculations using the Bernoulli equation may need to be considered as a starting point to a more detailed analysis (which is not required in the IB Physics course).

The Bernoulli effect

Revised

- The Bernoulli equation shows us that if the speed of a fluid increases, it will exert less pressure. Three applications of this important effect are given below.

Explaining situations involving the Bernoulli effect

1 **Aerofoils (airfoils).** The cross-sectional shape of an object can be designed to produce a force when a fluid flows around it. Figure 14.43 shows the streamlines around an aircraft wing. The shape of the wing causes the air to flow faster above the wing and therefore exert a lower pressure than the slower moving air beneath the wing. An inverted aerofoil on the back of a car can increase the downwards force, thereby increasing friction between the tyres and the road.

2 **Spinning balls.** See Figure 14.44. A ball can be made to curve in flight by making it spin.

3 **Venturi tubes.** The drop in pressure that occurs when a fluid flows through a constriction in a tube can be used to force a second fluid to mix with the fluid in the pipe (see Figure 14.45). A typical use of this effect is the mixing of fuel and air in an engine.

Key concept

The reduction in pressure when fluid speed increases is generally known as the **Bernoulli effect** and it has many applications.

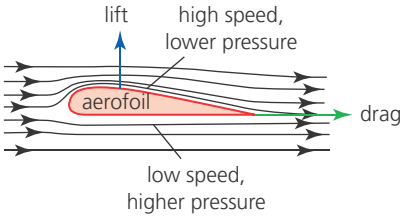


Figure 14.43

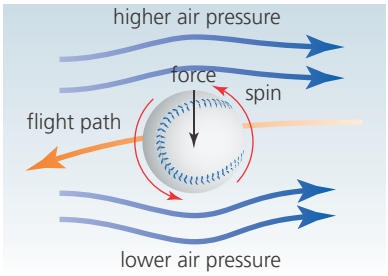


Figure 14.44

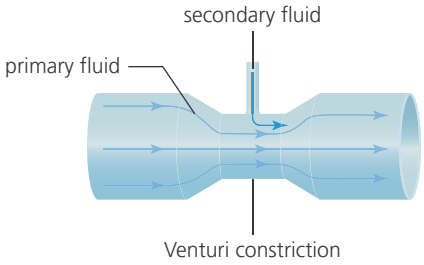


Figure 14.45 Venturi tube.

QUESTIONS TO CHECK UNDERSTANDING

68 Use the Bernoulli effect to explain why the ball shown in Figure 14.44 can be made to change direction.

69 Use a diagram to show how wind passing around the outstretched sail on a boat can cause a force helping to push the boat forward.

Viscosity

Revised

- Viscosity is a measure of a fluid's resistance to flow. Earlier in this section we noted that *ideal fluids* have zero viscosity, but this is not realistic.
- The streamlines in Figure 14.46 represent the flow of a fluid with significant viscosity. The velocity of flow varies from zero at the walls to a maximum in the centre of the container.
- Typically, the viscosity of a liquid varies significantly with changes in temperature. (For example, the viscosity of water at 20°C is 1.0×10^{-3} Pas, while the viscosity of a type of engine oil at the same temperature may be 300× greater at about 0.3 Pas, but at a temperature of 100°C the viscosity of the same oil has decreased significantly to about 10×10^{-3} Pas, whereas the viscosity of water has fallen to 0.3×10^{-3} Pas.)

Key concept

Viscosity, η, can be considered as a measure of a fluid's internal friction between streamlines. Viscosity has the SI unit Pas.

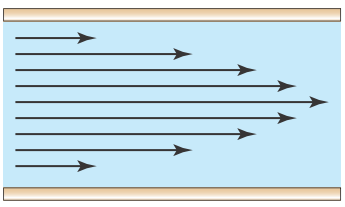


Figure 14.46

Describing the frictional drag exerted on small spherical objects in laminar fluid flow

- We saw in Chapter 2 that when objects move through fluids they experience **viscous drag forces**. In general terms, the magnitude of the drag force, for a given fluid, was considered to depend on the shape of the object, its cross-sectional area and its speed.

- The magnitude of the drag force also depends on the viscosity of the fluid.
- Similar comments apply to the force exerted on a stationary object by a moving fluid.
- To understand this in more detail, we will simplify the situation to consider the motion of a small spherical object (with a smooth surface) through a liquid with laminar flow.

Stokes' law and viscosity

- **Stokes' 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' law is often used to determine viscosity using data from experiments involving the terminal speed of falling spheres. Figure 14.47 shows the forces acting on such an object and the streamlines around it.
- When the sphere has reached its terminal speed, v_t (Chapter 2), there is no resultant force.
- Viscous drag, F_D + buoyancy force, B = weight, or $\boldsymbol{6\pi\eta rv_t + \rho_f V_f g = mg}$

Key concept

The viscous drag acting on a smooth sphere can be determined from $\boldsymbol{F_D = 6\pi\eta rv}$, but this equation (Stokes' law) only applies to laminar flow.

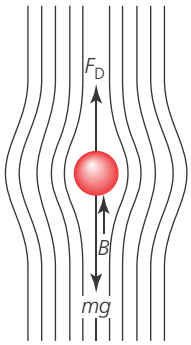


Figure 14.47

Solving problems involving Stokes' law

QUESTIONS TO CHECK UNDERSTANDING

70 a A sphere of radius 5.2 mm is moving horizontally through air with a speed of 8.3 m s^{-1}. If the viscosity of air is 1.8×10^{-5} Pa s, determine the viscous drag force on the sphere.

b What assumptions did you make when answering (a)?

c Calculate the instantaneous deceleration produced on the sphere if its mass was 4.8 g.

71 A small sphere of radius 3.5 mm and mass 2.7 g was dropped through oil of viscosity 0.24 Pa s and density 880 kg m^{-3}. Determine its maximum speed.

Laminar and turbulent flow and the Reynolds number

Revised

- At low flow speeds, most fluids exhibit laminar flow, but if the speed of flow of a fluid is gradually increased, this regular flow will become disrupted and turbulent flow will begin (see Figure 14.48, which compares the two types of flow).
- In **turbulent flow**, instead of layers, there are unpredictable eddies, vortices and other irregular motions. Figure 14.49 is a photograph of such turbulence around a square object in a wind tunnel.
- The speed at which turbulent flow begins depends on the density and viscosity of the fluid and the dimensions of the system containing the fluid.
- 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 *Reynolds number, R*.

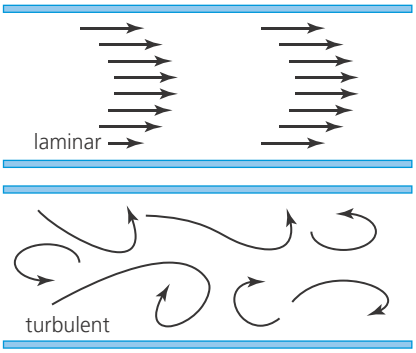


Figure 14.48

Key concept

If the speed of a fluid is increased enough, its flow changes from laminar to turbulent.

Reynolds number, R can be used to predict the flow speed, v, at which this may occur in a tube: $\boldsymbol{R = \frac{vr\rho}{\eta}}$ (r is the diameter of the pipe).

- Turbulence by its nature is unpredictable, but it is likely to occur if R exceeds a certain value for any particular system. As a generalized guide, *if $R > 1000$, we can expect turbulent flow.*

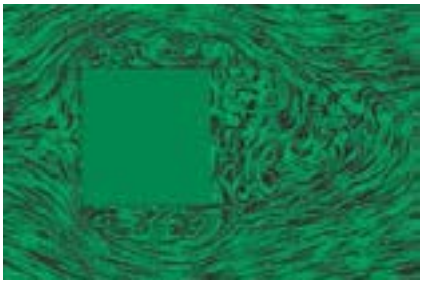


Figure 14.49

Determining the Reynolds number in simple situations

- As an example, consider the flow of oil of density $9.3 \times 10^2\,kg\,m^{-3}$ and viscosity 0.22 Pa s. If this oil flows through a pipe of radius 8.0 cm, using the above equation predicts that turbulence will occur if the oil flows faster than about $31.5\,m\,s^{-1}$.

Key concept

Turbulant flow can be expected if $R > 1000$.

QUESTIONS TO CHECK UNDERSTANDING

72 a Calculate the radius of the tube in the previous example which would result in the onset of turbulence at an oil speed of $1.0\,m\,s^{-1}$.

b If the oil was at a lower temperature, how would this affect your answer to (a)?

73 Air in an air conditioning system flows through a pipe of radius 5 cm. If the density of air is $1.22\,kg\,m^{-3}$ and its viscosity is 1.8×10^{-5} Pa s, determine whether an air speed of $1.0\,m\,s^{-1}$ will be laminar or turbulent.

NATURE OF SCIENCE

Human understanding

The properties of fluids in motion cannot be explained by an understanding of the behaviour of the individual particles that they contain. For example, the model of an ideal gas (Chapter 3) cannot fully explain the characteristics of a gas flowing through a pipe. New models have to be developed for the flow of fluids and these could then be used to explain the flow of air around vehicles and the flow of fluids through pipes.

14.4 Forced vibrations and resonance (Additional higher level)

Revised

Essential idea: In the real world, damping occurs in oscillators and has implications that need to be considered.

Natural frequency of vibration

Revised

- After they have been disturbed, many objects and structures have a *natural frequency* (or frequencies) at which they will oscillate. Striking a drum is a simple example. The oscillations of a vibrating hacksaw blade are more easily observed, and these are commonly used in school investigations (see Figure 14.50).

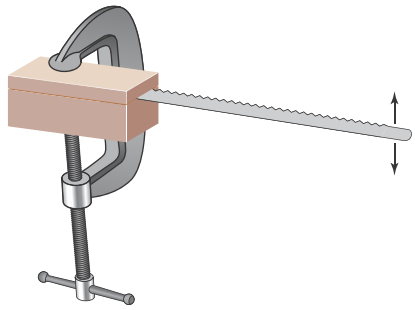


Figure 14.50

Key concept

The **natural frequency** of an object or system is the frequency with which it oscillates when there are no external periodic forces acting on it.

Energy of an oscillator

Revised

- Remember from Chapter 4 that there is a continuous exchange between potential energy and kinetic energy in any mechanical oscillator.

Key concept

The total energy of an oscillator is proportional to its amplitude squared.

Q factor and damping

Revised

- Frictional forces dissipate energy and so reduce the speeds and amplitudes of oscillations. So that successive amplitudes, A, of a system oscillating naturally will get less and less as shown in Figure 14.51, in which $A_1 > A_2 > A_3$, etc.

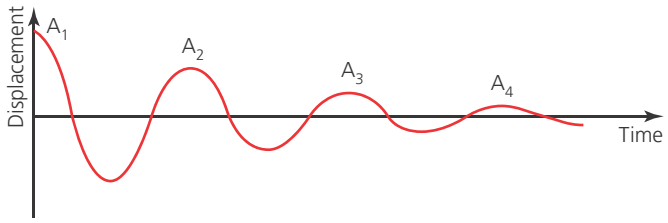


Figure 14.51

- The ratio of successive amplitudes may be considered to be constant, $\frac{A_1}{A_2} = \frac{A_2}{A_3} = \frac{A_3}{A_4}$, etc., so that the decrease in amplitude with time can be described as exponential.

QUESTIONS TO CHECK UNDERSTANDING

74 a What factors affect the natural frequency of oscillation of:

i a simple pendulum

ii a mass hanging on a spring?

b Suggest three factors which affect the natural frequency of a hacksaw blade oscillator, such as shown in Figure 14.50.

75 Successive amplitudes of a simple pendulum were measured to be 5.8 cm, 5.1 cm and 4.5 cm. Predict the amplitude of the next oscillation.

- The damping in an oscillator is represented numerically by its Q *(quality) factor*. An oscillator which experiences a lot of damping has a *low* Q (quality) factor.
- For example, suppose in a damping experiment successive amplitudes were measured to be 12, 7, 4 and 2 cm (similar results to those shown in Figure 14.51). A mathematical check will confirm that the ratios of successive values are all about 0.60, suggesting an exponential decrease. The corresponding energies (proportional to amplitude *squared*) have ratios of about $0.60^2 = 0.36$. Q factor $= 2\pi \times \left(\frac{1.0}{0.36}\right) \approx 6\pi$.

Qualitatively and quantitatively describing examples of under-, over- and critically damped oscillations

- Damping can occur to very different *degrees* in different systems. It may be desirable or unwanted. It is convenient to identify three special cases, these are shown in Figure 14.52.
- When an oscillator returns relatively quickly towards its equilibrium position without oscillating, it is described as being *critically damped*. The Q factor for critical damping is usually quoted to be about 0.5.
- A system taking much longer to return to its equilibrium position is described as **over-damped**.
- In comparison, an **under-damped** system dissipates less energy and therefore has a larger Q factor, so that oscillations (maybe many) will occur.

Key concept

The dissipation of useful energy from an oscillator because of resistive forces is called **damping**.

Damping usually reduces successive amplitudes exponentially.

Common mistakes

Under most circumstances, we can assume that the dissipation of energy from an oscillator does *not* affect its period. Distances, speeds and acceleration all decrease, but each oscillation still takes the same amount of time.

Key concept

The **Q (quality) factor** of an oscillator is a way of representing the amount of damping involved:

$$Q = 2\pi\left(\frac{\text{energy stored}}{\text{energy dissipated per cycle}}\right).$$

Key concept

Critical damping occurs when a system returns relatively quickly towards its equilibrium position without passing through it.

In comparison, other oscillators may be described as *over-damped* or *under-damped*.

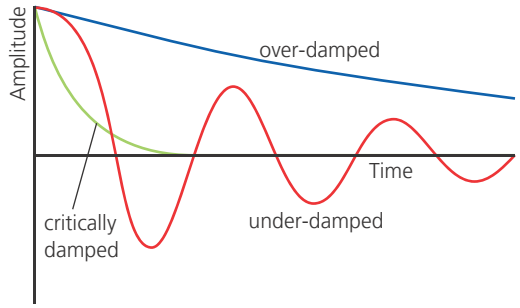


Figure 14.52

Solving problems involving Q factor

QUESTIONS TO CHECK UNDERSTANDING

76 Consider the oscillator shown in Figure 14.50.

- **a** What degree of damping would you expect to observe?
- **b** Suggest how you could decrease its Q factor.

77 Determine the Q factor for the pendulum discussed in Question 75.

78 Doors that can swing both ways are common in hospitals and restaurants (see Figure 14.53). Describe how these doors would behave if they were

- **a** critically damped
- **b** under-damped
- **c** over-damped.

79 The elastic strain energy stored in a stretched spring $= \frac{1}{2}kx^2$ (see Chapter 2).

- **a** Explain the meaning of the symbols x and k.
- **b** A spring which is stretched by hanging a mass on its end is then made to oscillate vertically. Determine a value for the Q factor of this system if the amplitude decreases from 5.7 cm to 5.3 cm in successive oscillations.

Figure 14.53

Periodic stimulus and driving frequency

Revised

Objects and systems are commonly exposed to **periodic stimulus** (forces) from external vibrations. We need to carefully distinguish between the frequency of any *external* vibration and the *natural* frequency with which the system would vibrate.

Key concept

If an external oscillation stimulates a system to oscillate, we refer to them as **forced oscillations**.

Expert tip

The concept of *oscillations* has been discussed extensively in earlier chapters. In this topic, the term *vibration* is also commonly used. Vibration is a term that can be used for an oscillation in any system which involves moving mass (especially if the frequency is relatively high). That is, a vibration is an oscillation in a mechanical system. For example, alternating currents or light waves may be described as oscillations, but not vibrations.

Resonance

Revised

- The effect of an external, *driving frequency* on a system depends on how it compares to the natural frequency.

Describing the phase relationship between driving frequency and forced oscillations

- A widely used and easily understood example of resonance is a child on a swing as shown in Figure 14.54. If the driving frequency (provided by the man pushing the swing) is the same as the natural frequency at which the swing oscillates without being pushed, then the amplitude will increase, *but only if the two oscillations are in phase with each other*. For example, if the driving force had the same frequency as the natural frequency, but was half an oscillation (π rad) *out of phase*, then the amplitude would be reduced.
- If the driving frequency is half the natural frequency (or one third, or one quarter, etc.), then some resonance effects will still be observed.

Key concept

When the **driving frequency** is similar to the *natural* frequency of a system (and is in phase with it), energy is efficiently transferred to the system and the amplitude of oscillation increases. This effect is called **resonance**.

Figure 14.54

Graphically describing the variation of amplitude of vibration with driving frequency of an object close to its natural frequency of vibration

- The degree of damping (Q factor) determines how significant the resonance effects are. In practice, the peaks shown in Figure 14.55 may shift to slightly lower frequencies with increased damping.

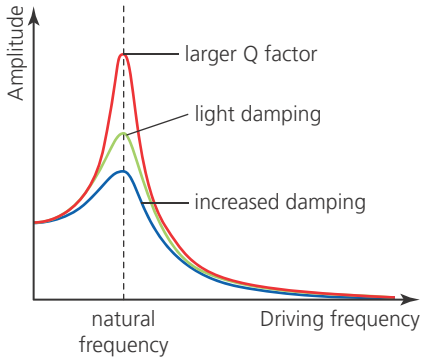


Figure 14.55

Key concept

A **frequency response graph** shows how amplitude can increase when the driving frequency is the same as, or is close to, the natural frequency. The same graph can be used to show the effects of damping (see Figure 14.55).

- *For a resonating system which is oscillating steadily*, power input = power loss = energy dissipated per cycle/period = energy dissipated per cycle × frequency.
- Using the equation given in the Key concepts box, we can determine that an oscillating system resonating with a frequency of 20 Hz, energy of 4.5×10^{-2} J and Q factor of 5 would need a continuous average power input of about 1 W.

Key concept

The Q factor of an oscillator can be related to its resonant frequency by the equation: $Q = 2\pi \times \text{resonant frequency} \times \left(\frac{\text{energy stored}}{\text{power loss}}\right)$.

Describing the useful and destructive effects of resonance

- Resonance can be *beneficial*. The absorption of infrared radiation by gases in the Earth's atmosphere is an important example which was discussed in Section 8.2. Other examples of useful resonance include the tuning of electronic circuits to receive certain transmitted frequencies and amplifying the intensity of musical sounds from stringed instruments by mounting them on boxes.

Expert tip

Another useful example of resonance is NMR: nuclear magnetic resonance, which is used for obtaining images of soft tissues inside the human body. Protons in the tissue are made to resonate at the same radio frequency as provided by electromagnetic waves generated by currents in coils around the patient.

- The flow of air or water around structures (buildings or bridges, for example) can produce oscillating forces which can cause damage, especially if resonance occurs.
- Earthquakes produce both transverse and longitudinal waves which will be particularly destructive for structures which have parts which have natural frequencies similar to those of the earthquake (see Figure 14.56).
- Engines and machines have oscillating parts. The vibrations from these can cause resonance problems in other parts of the device.
- Structures can be designed to limit the destructive effects of resonance. This can be achieved by changing the masses and the stiffness of parts of the system. From Chapter 9, Section 9.1, we have a simplified model: the natural frequency of oscillation of a mass on a spring is dependent on the magnitude of the mass and the force constant of the spring. A way of dissipating energy safely can also be incorporated into the design.

Figure 14.56

QUESTIONS TO CHECK UNDERSTANDING

80 Figure 14.57 shows an experiment to investigate resonance. The maximum amplitude of oscillation of the mass is measured for different frequencies applied to the vibration generator. The mass was 300 g and the force constant of the spring was 7.8 N m^{-1}.

- **a** Use information from Section 9.1 to determine the natural frequency of vibration of the mass on the spring.
- **b** Sketch an amplitude-frequency graph to represent the results you would expect from this experiment.
- **c** Show how the results would change if the mass was surrounded by water.

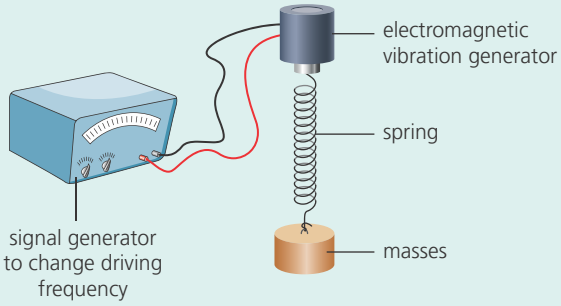


Figure 14.57

81 Calculate the average power loss from a system resonating at 4.6 Hz, if it had a total energy of 740 mW and a Q factor of 95.

82 Confirm that an oscillating system resonating with a frequency of 20 Hz, energy of 4.5×10^{-2} J and average power input of 0.95 W has a Q factor of about 6.

83 Use the concept of resonance to explain the greenhouse effect.

84 Discuss how the concept of resonance may be needed to explain why some buildings suffer worse damage in an earthquake than others.

NATURE OF SCIENCE

Risk assessment

The potentially destructive effects of resonance are more common than may at first be apparent. Every structure, large or small, may be damaged by vibrations from its surroundings, whether those vibrations are naturally occurring or produced by machinery. Designers, architects and engineers need to consider the potential risks posed by vibrations and resonance on their structures.

Option C 15 Imaging

15.1 Introduction to imaging

Revised

Essential idea: 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.

Thin lenses

Revised

- A **lens** is a piece of glass or plastic with regularly curved surfaces which uses the phenomenon of *refraction* to converge or diverge light in a predictable way.
- A material is said to be **transparent** if light is transmitted and objects can be seen through it. A material is described as **opaque** if light cannot be transmitted through it.
- *Thin lenses* which are not too large in diameter behave in the most predictable ways. Thin lenses are the only lenses discussed in the IB Physics course.

Describing how a curved transparent interface modifies the shape of an incident wavefront

- Light can be considered to travel as waves and we saw in Chapter 4 that the movement of waves can be modelled by drawing *wavefronts*.
- When wavefronts cross interfaces between two media (like air/glass), their speed changes and this will result in changes of direction (unless their motion is perpendicular to an interface). This effect is called *refraction* and it was explained in Chapter 4. If the interfaces have regular curvatures, then the resulting wavefronts can be converged or diverged.
- Figure 15.1 shows wavefronts travelling from a point on the left, moving more slowly through a lens, and then being converged to a point on the right, called a **focus**.

Key concept

When wavefronts pass through a regularly curved interface between two transparent media, refraction changes their shape and can make them converge towards a point, or diverge away from a point.

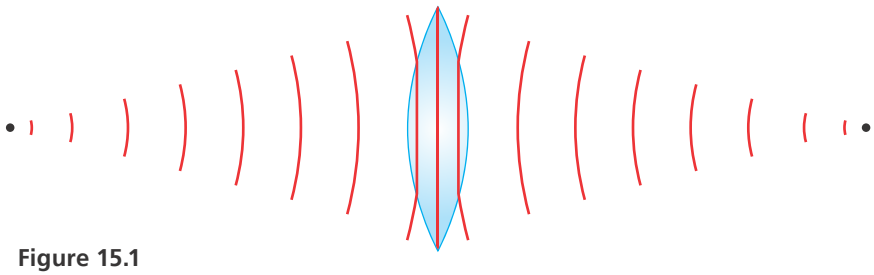


Figure 15.1

Ray diagrams

Revised

- Diagrams like Figure 15.1 can be difficult to draw. The effects of lenses on light are usually much better represented by using **ray diagrams**. Reminder from Chapter 4: *rays* are lines which show the direction in which wavefronts are travelling. Rays are always perpendicular to wavefronts.
- Figure 15.2a is a ray diagram representing the same situation as shown by the wavefronts in Figure 15.1. We are not usually concerned with exactly how a ray passes through a lens itself, so it is common to show rays changing direction only in the middle of a lens. This is shown in Figure 15.2b, which simplifies the diagram even further by using a line to represent the lens.

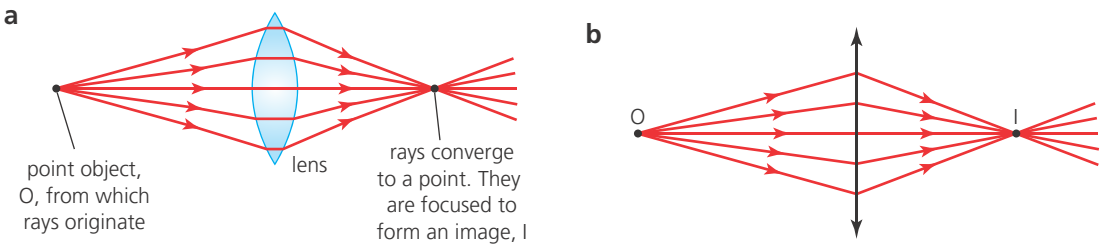


Figure 15.2

- A source of light rays is usually called an **object**, and an **image** (of the object) is formed where the rays focus. For simplicity, Figure 15.2 shows *point* objects and images, but in practice they will usually be larger (*extended*).
- Topics within the study of optics in which the wave properties of light are generally considered insignificant, so that light is shown travelling in straight lines, are called *geometric optics*.

Converging and diverging lenses

- When light rays are incident on a lens that is thicker in the middle than at its edges, the rays will usually be converged to form an image at the point where the rays are focused, as shown again in Figure 15.2a, but this time the incident rays and wavefronts are parallel.
- If rays are incident on a lens that is thinner in the middle than at its edges, they will be diverged. This is shown in Figure 15.3b. Diverging rays form virtual images (see later).

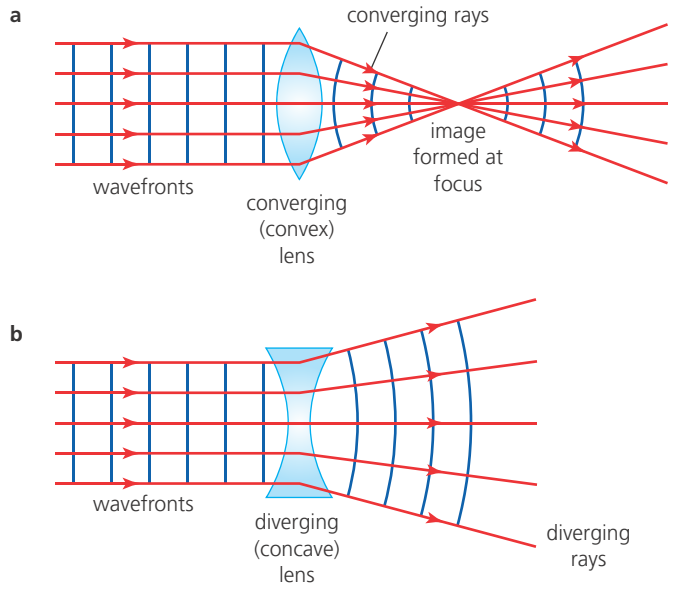


Figure 15.3

Identifying the principal axis, focal point and focal length of a simple converging or diverging lens on a scaled diagram

- The *principal axis* is a key feature of any ray diagram.
- A lens has two focal points, both the same distance from the lens. (The focal point is sometimes called the principal focus.)
- The *focal length* of a lens depends on the refractive index of the material and the curvature of its surfaces.

Key concept

The reflection and refraction of wavefronts is easier to understand when we use straight rays in diagrams to represent the direction in which the wavefronts are travelling.

Key concepts

A lens which is thicker in the middle than at its edges will converge the rays that are passing through it (unless they come from an object which is too close to the lens). Such lenses are called **converging lenses**.

Diverging lenses are thinner in the middle than at the edges and they cause rays to diverge.

Key concepts

The **principal axis** of a lens is the imaginary straight line passing through the centre of the lens, which is perpendicular to its surfaces.

The **focal point** of a lens is 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 the distance between the centre of the lens and the focal point.

- The meanings of these terms are shown in Figure 15.4 for both types of lens.

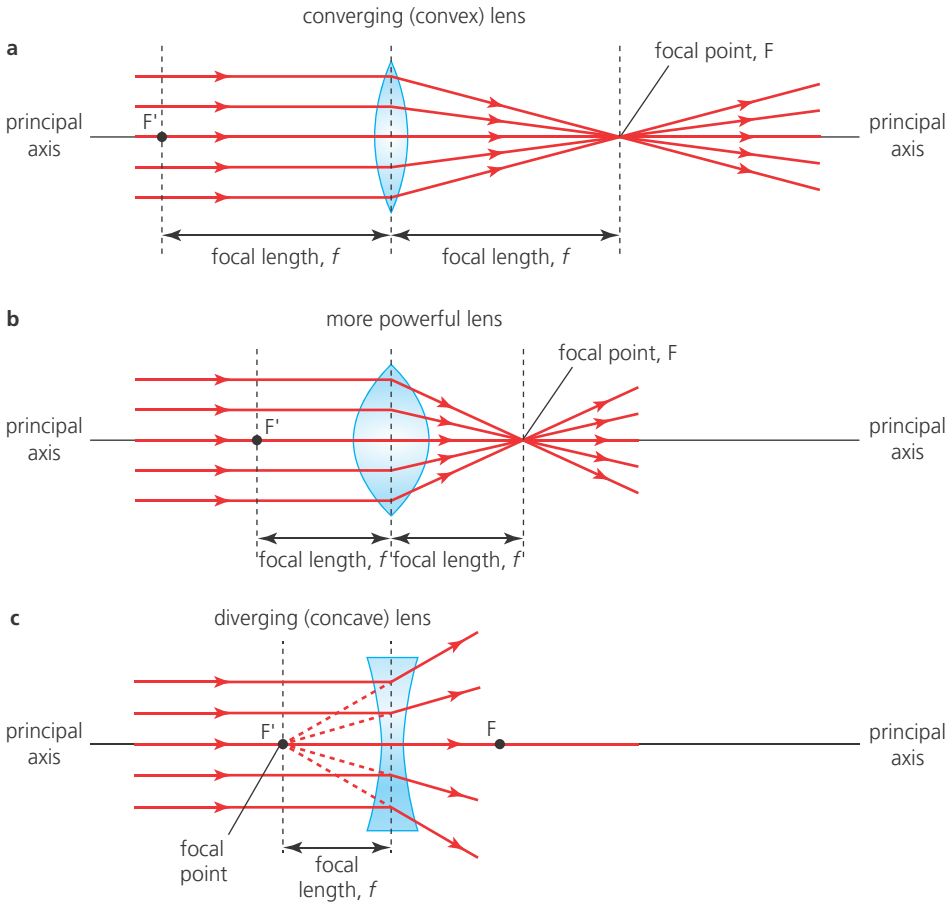


Figure 15.4

- Comparing Figure 15.4a with Figure 15.4b, the effect of lens shape on focal length can be seen. Greater surface curvature produces a shorter focal length, and the lens is described as being more *powerful*.
- If the focal length of a lens is given in metres, the calculated power has the unit of **dioptres**, D. For example, a converging lens of focal length 50 cm has $P = \frac{1}{0.50} = 2.0\,\text{D}$. As we shall see, the focal lengths of diverging lenses are given *negative* values, so that their powers are negative too. For example, a lens with a power of –5.0 D is a diverging lens of focal length 20 cm.

Real and virtual images

- All images can be described as being *real* or *virtual*. Real images are seen where light rays converge (usually on a screen). Virtual images are seen by looking through a lens or at a mirror.

QUESTIONS TO CHECK UNDERSTANDING

1 Consider Figure 15.3a.
   - a What is the shape of the wavefront inside the lens (which is not shown in the diagram)?
   - b How would the diagram change if the lens was changed for another which had the same shape but was made with glass of a greater refractive index?

2 a Draw a ray diagram to represent light rays from a point object passing into and out of a plano-diverging lens. (This type of lens has one surface which is flat.) Label the principal axis and focal points.
   - b Draw a second diagram to show the effect of changing the lens for another which had a surface of greater curvature.

Key concept

The (optical) **power of a lens** is defined as $\frac{1}{\text{focal length}}$; $P = \frac{1}{f}$.

Expert tip

As people get older and their eye muscles get weaker, they often need the help of a converging lens in order to see things close to them. Glasses (spectacles) for reading typically have a power of +2 D or +3 D. Conversely, the focusing in the eyes of some younger people is too powerful so that they need diverging lenses to help them see better. A typical power for a lens to correct short-sight is –5 D.

Key concept

Real images are formed where real rays actually converge to a focus. **Virtual images** can only be seen by looking through lenses, or mirrors, to the points from which the rays appear to diverge.

3 a What is the power of a converging lens which has a focal length of 25 cm?

b A lens has a power of –1.5 D.

i What type of lens is it?

ii What is the focal length (cm) of the lens?

4 a What is the essential difference between the kind of image seen on a screen at a cinema and an image seen in a bathroom mirror?

b Are the images produced in:

i an eye, real or virtual?

ii a camera, real or virtual?

Using ray diagrams to predict the properties of images

Revised

- We can fully describe any image by its position, size, whether it is upright or inverted, and whether it is real or virtual.
- These properties can change when the distance between an object and a lens is varied.
- The position and nature of an image can be determined by using a scaled ray diagram, in which the paths of three rays are always known, as shown in blue in Figure 15.5.
- In ray diagrams, we assume that we are representing a *thin lens* and the rays are close to the principal axis (even if this is not well represented in the drawing!) If this is not true, the image will not be formed exactly where predicted and the focus/image will not be as well defined.

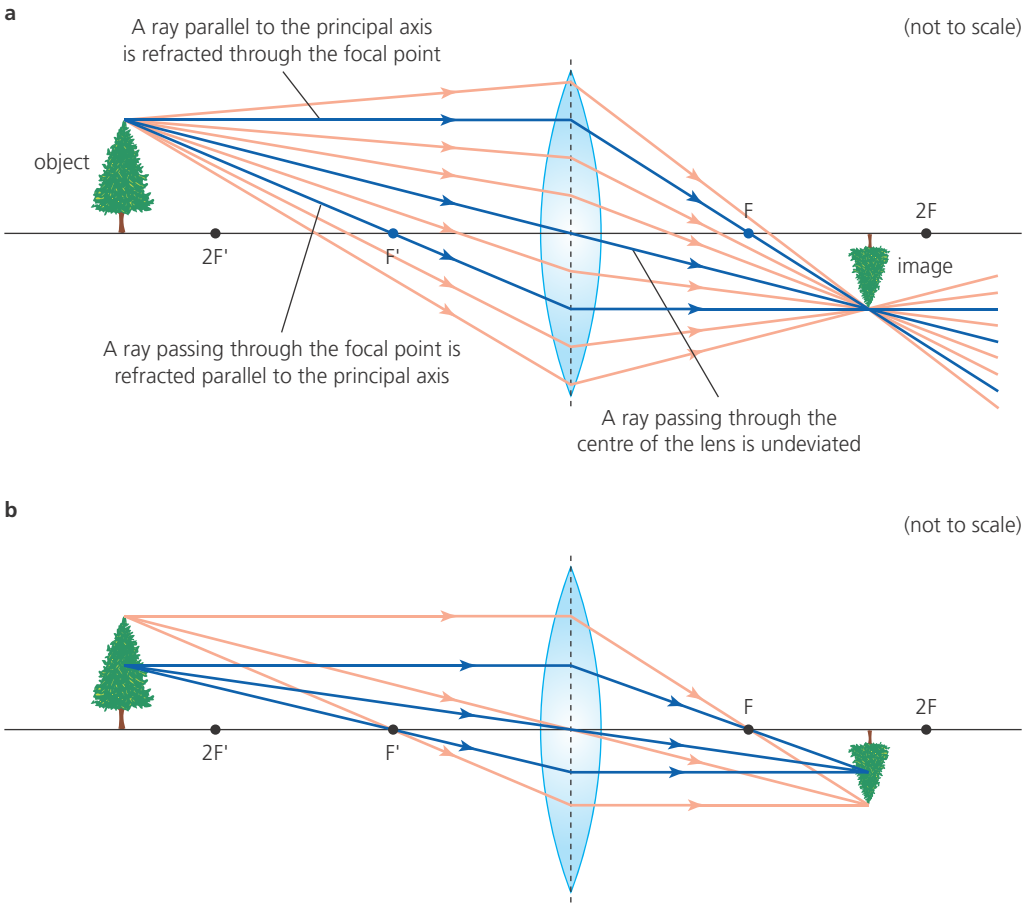


Figure 15.5

Converging lenses

- The ray diagrams in Figure 15.6a show how the real image changes as the distance between an object and a converging lens is varied.
- Figure 15.6b shows the formation of a magnified virtual image by a single converging lens placed close to an object (closer than F). Used in this way it is described as a *simple magnifying glass* (discussed in more detail later).

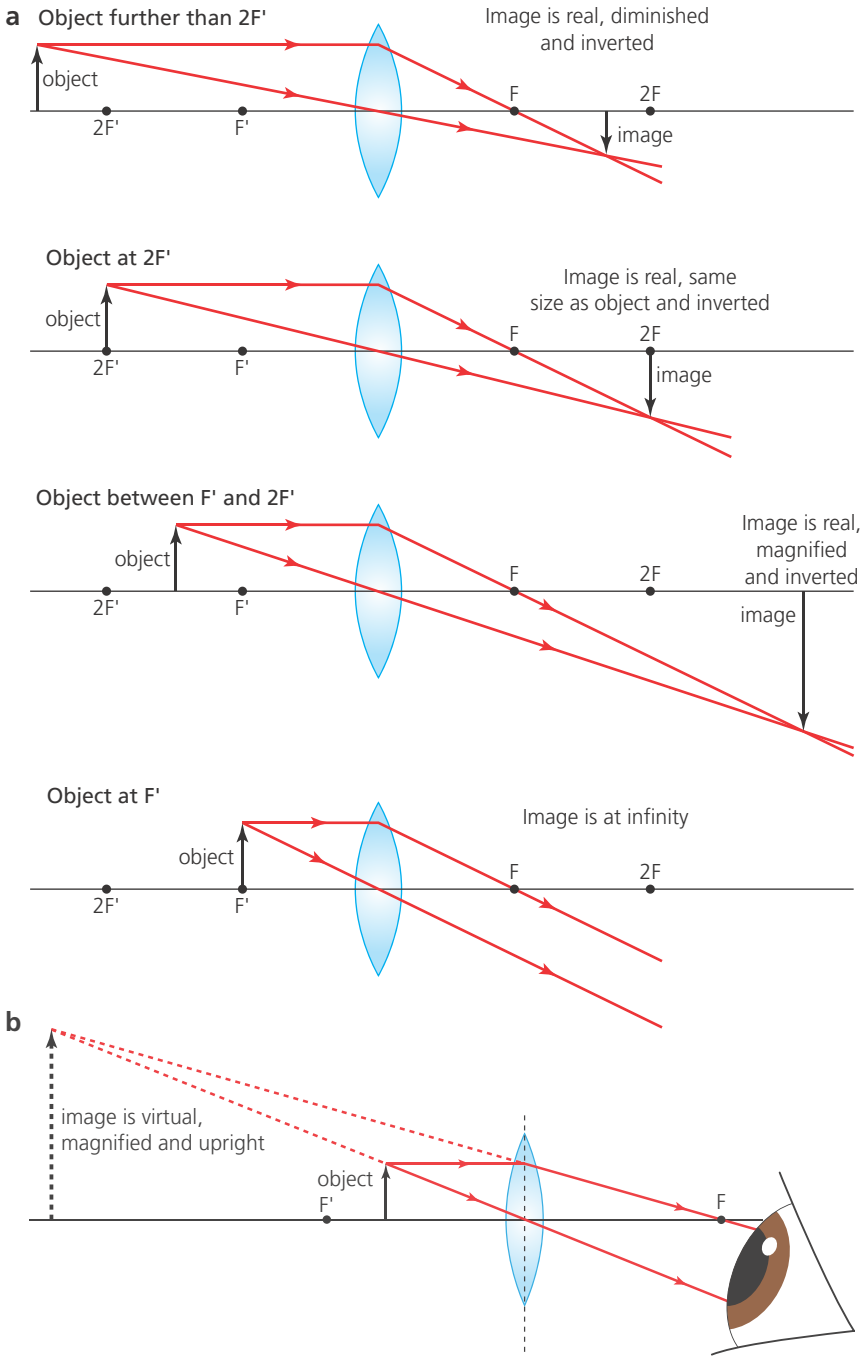


Figure 15.6

Key concepts

Ray diagrams drawn to scale can be used to predict the properties of the images formed by objects placed in different positions.

If an object is a long way from a *converging lens*, a smaller, real, inverted image is formed on a screen placed close to the focal point. As the object and lens get closer together, the image moves further away from the lens and gets larger (and dimmer), but it remains real and inverted.

If the object is closer to the lens than the focal point, no real image can be formed, but a magnified, upright, virtual image can be seen by looking through the lens.

Diverging lenses

- Figure 15.7 shows a typical ray diagram for image formation by a diverging lens.

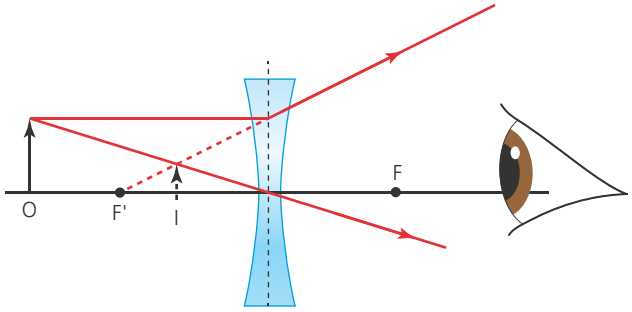


Figure 15.7

Key concept

Ray diagrams for diverging lenses confirm that the images are always upright, virtual and diminished.

Solving problems involving not more than two lenses by constructing scaled ray diagrams

- Figure 15.8 shows a ray diagram for the formation of an image using two lenses. The blue lines are construction lines used to locate the top of the final image.

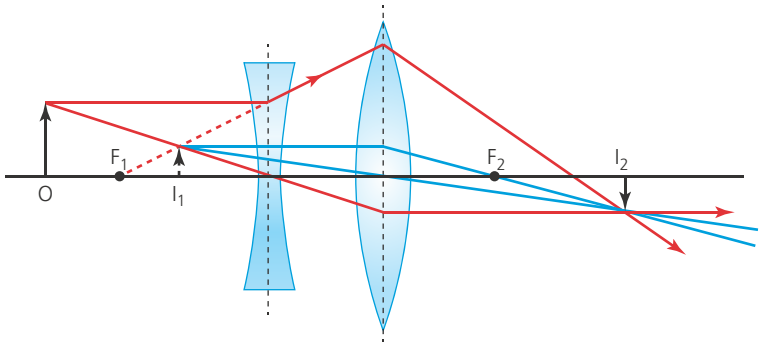


Figure 15.8

Key concept

Ray diagrams can also be used to locate the image formed by a system of *two lenses*. The image formed by the first lens is treated as the object for the second lens.

QUESTIONS TO CHECK UNDERSTANDING

5 Draw a ray diagram to determine the nature and position of the image formed when an object is placed 25 cm from a converging lens of focal length 10 cm.

6 An upright, virtual and diminished image of height 1.0 cm is formed 4.0 cm from a diverging lens of focal length 6.0 cm. Determine the location and height of the object by drawing a scaled ray diagram.

7 A student uses a converging lens to project a magnified image on to a screen which is 2.0 m away from the object. Use a ray diagram to determine the position and focal length of the lens if the image is ten times bigger than the object.

8 Two converging lenses, both having a focal length of 5.0 cm, are placed 20.0 cm apart. If an object of height 2.0 cm is placed 8.0 cm in front of the first lens, use a ray diagram to determine the position and nature of the final image after the light has passed through the second lens.

Using the thin lens equation to predict the properties of images

Revised

- u is the symbol used for *object distance* and v is the symbol used for *image distance*, as shown in Figure 15.9, which also shows the heights of the object and image.
- It should be remembered that this equation only applies for thin lenses with rays close to the principal axis.

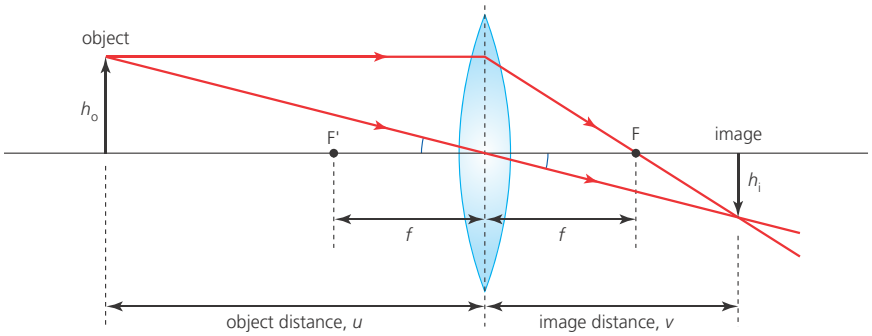


Figure 15.9

Key concept

As an alternative to a ray diagram, the thin lens equation can be used to determine the position of an image

$$\frac{1}{f}=\frac{1}{v}+\frac{1}{u}$$

When using this equation, it is important to remember that the focal length of a diverging lens and the distance to a virtual image are always negative.

'Real-is-positive' convention

- It is possible to put values for f and u (when $u<f$) into the thin lens equation 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.

Common mistake

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:

- *Positive*: focal lengths of converging lenses, distances to real objects and images, magnification of upright images.
- *Negative*: focal lengths of diverging lenses, distances to virtual images, magnification of inverted images.

Range of human vision

- The nearest point to the human eye at which an object can be clearly focused (without straining) is called the **near point** of the eye.
- The furthest point from the human eye that an object can be clearly focused is called the **far point** of the eye. For a normal eye, this is assumed to be at infinity.

Key concept

The distance from the eye to the near point is given the symbol D and it is assumed to be 25 cm for a normal eye.

Expert tip

The distance between the lens and the retina (where real images are formed) in an adult human eye is about 2.5 cm. This must also be the focal length of the eye's focusing system, so that parallel rays from distant objects are correctly focused. The shape of the lens is changed by optic muscles to adjust the focal length in order to correctly focus objects at different distances away. If eyes are not capable of doing this well, then extra lenses (spectacles or contact lenses) can be worn.

Linear and angular magnification

Revised

- The **linear magnification**, m, of an image is the ratio of the height of the image, h_i, divided by the height of the object, h_o (any other linear dimension can be used instead of height): $m = \frac{h_i}{h_o}$. Magnification is a ratio, so that is has no units.
- A magnification of 1 means that the object and image are the same size. A magnification of less than one means that the image is *diminished*: i.e. it is smaller than the object.
- From Figure 15.9 we can use geometry to show that $m = \frac{h_i}{h_o} = -\frac{v}{u}$. The negative sign has been added in this equation because of the 'real-is-positive' convention (see page 76).
- Sometimes the dimensions of an object and/or an image are not easily determined, for example, an object may be a very long way away, or an image may be virtual. Sometimes quoting a value for a linear magnification may be misleading, for example the linear magnification of an impressive 1 m diameter image of the Moon would be very small. Under circumstances like these *angular magnification* is more useful.
- **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. $M = \frac{\theta_i}{\theta_o}$. This is illustrated in Figure 15.10.

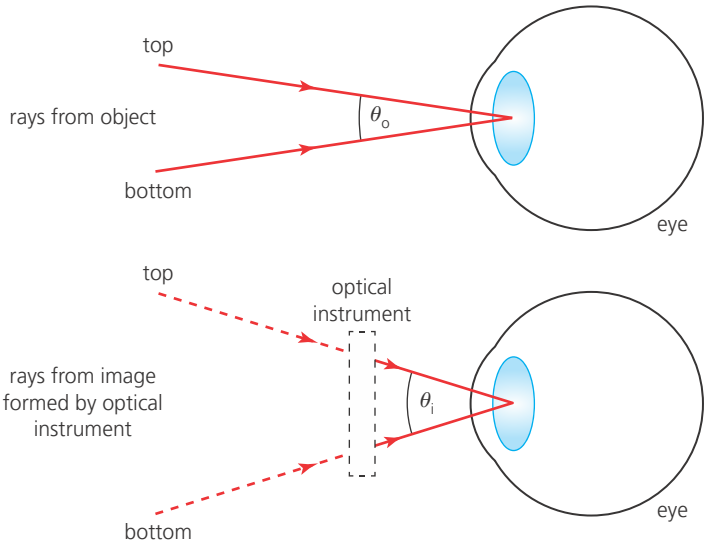


Figure 15.10

Key concepts

The magnification of an image may relate to the linear sizes of an object and image, or to the angles they subtend at the eye.

Linear magnification, $m = \frac{h_i}{h_o} = -\frac{v}{u}$.

(The inclusion of a negative sign means that virtual upright images have a positive magnification and real inverted images have a negative magnification.)

Angular magnification, $M = \frac{\theta_i}{\theta_o}$.

Magnification produced by a simple magnifying glass

- Figure 15.11 shows a magnifying glass being used so that a clear image is formed as close as possible at the *near point of the eye*. In this position, the magnification is the greatest possible.
- The *linear magnification*, $m = \frac{h_i}{h_o}$ (as above), but using geometry, we can show that the more useful *angular magnification*, $M_{\text{near point}} = \frac{D}{f} + 1$.

Key concepts

If an object is placed closer to a converging lens than its focal point, the lens will act as a **simple magnifying glass**.

If the lens is moved to obtain the largest clear image, the image will be formed at the near point and the angular magnification can be determined from

$M_{\text{near point}} = \frac{D}{f} + 1$

The eye can be more relaxed if the lens is moved so that the image is at infinity, but the angular magnification is then reduced:

$M_{\text{infinity}} = \frac{D}{f}$

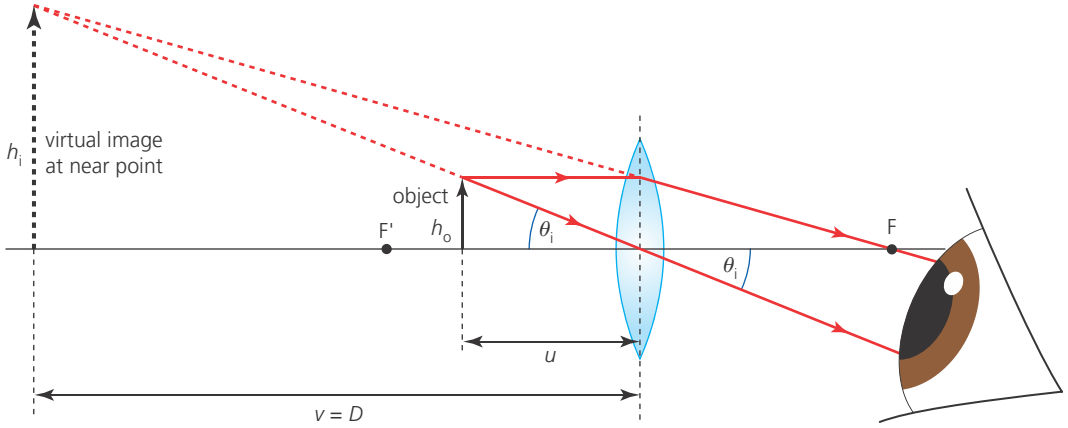


Figure 15.11

Solving problems involving the thin lens equation, linear magnification and angular magnification

QUESTIONS TO CHECK UNDERSTANDING

9 **a** Use the thin lens equation to determine the position of the image formed of a 3.2 mm high object placed 24 cm in front of a lens of power +10 D, and whether it is real or virtual.

b Calculate the linear magnification of the image and whether it is upright or inverted.

10 A lens produced a magnification of –2.4.

a What kind of lens was this?

b If the object was 4.5 cm from the lens, where was the image?

c Use the thin lens equation to determine the focal length of the lens.

11 Consider again Figure 15.8. Use the thin lens equation to determine the position and magnification of the final image if the object is 25 cm from the diverging lens of focal length 16 cm, the converging lens has a focal length of 12 cm and the separation of the lenses is 14 cm.

12 Figure 15.12 shows a partial eclipse of the Sun.

a What angle does this image of the Sun subtend at your eye when it is at the near point of your eye?

b The Sun has a diameter of 1.4×10^6 km and it is 1.5×10^8 km from Earth. What is the angular magnification of Figure 15.12?

Figure 15.12

13 **a** Draw a ray diagram of a simple magnifying glass with the final image at infinity.

b Use your diagram to confirm that the angular magnification, $M_{infinity} = \frac{D}{f}$

14 A converging lens of focal length 10 cm is used as a magnifying glass, as shown in Figure 15.11.

a Where must an object be placed in order for the image to be formed at the near point of the eye?

b What is the linear magnification of this arrangement?

c What is the angular magnification of this arrangement?

Spherical and chromatic aberrations

Revised

- **Aberrations** are the (undesirable) properties of a lens (or mirror) which prevent the ideal situation in which rays from a point object produce a point image.
- Lens aberrations (especially with higher-power lenses) are the principal limitations on the magnification achievable by optical instruments that use lenses.

- In the IB Physics course, two particular kinds of aberration are discussed: (1) problems arising from the curvature of the lens surfaces and (2) problems arising from the fact that light usually has a range of different wavelengths.

Explaining spherical and chromatic aberrations and describing ways to reduce their effects on images

- *Spherical aberration* is illustrated in Figure 15.13 (in exaggerated form).

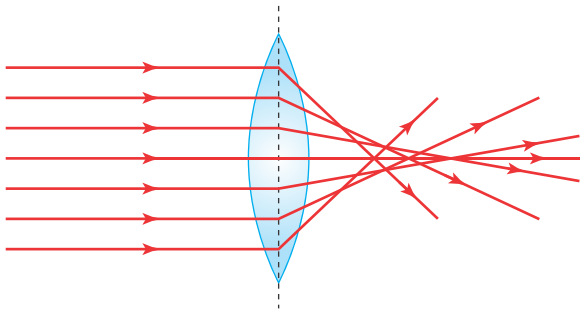


Figure 15.13

- The effect may be reduced by adapting the shape of the lens to become parabolic, and/or by only using the centre of the lens.
- *Chromatic aberration* occurs because refractive indices vary slightly with colour (wavelength) (see Figure 15.14). The effect can be reduced by combining lenses of different shapes and refractive indices, as shown in Figure 15.15.

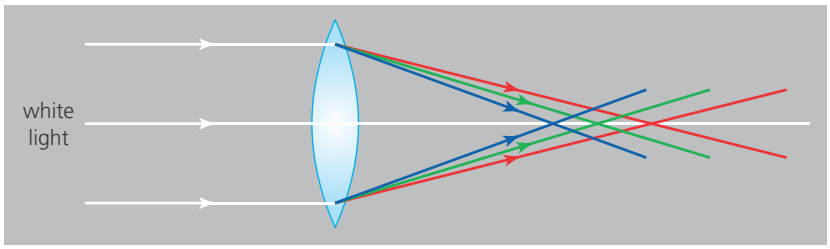


Figure 15.14

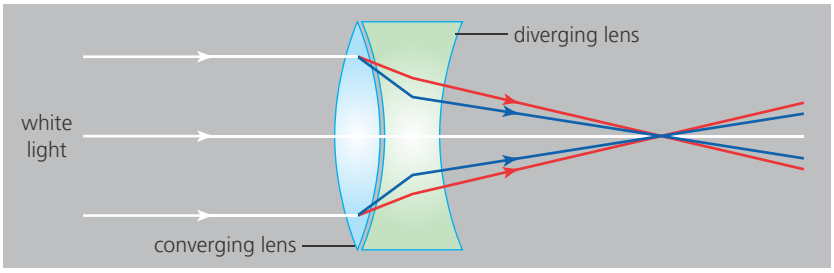


Figure 15.15

Key concepts

Spherical aberration is the inability of a lens or mirror which has spherically shaped surfaces to bring all (monochromatic) rays from the same point on an object to the same point focus.

Chromatic aberration is the inability of a lens to bring rays of different colours (from a point object) to the same focus.

QUESTIONS TO CHECK UNDERSTANDING

15 a Explain with the help of a diagram how the effects of spherical aberration can be reduced by only using the centre of a lens.

b Why would you expect spherical aberration to be worse with higher power lenses?

16 a Explain with the help of a diagram why images produced by single lenses tend to have coloured edges.

b What is the name of this effect?

Converging and diverging mirrors

Revised

- Mirrors with curved surfaces can also be used to focus images.
- Figure 15.16 shows an example: a **converging mirror** is producing a real, inverted, diminished image. The point C is the **centre of curvature** of the mirror surface. The focal point is half way between C and the mirror. A ray directed to, or from, the centre of curvature will reflect back off the mirror along the same path.
- In Figure 15.17, the object has been moved closer to the same mirror (closer than the focal point), so that the image is virtual, upright and magnified.
- Figure 15.18 shows the action of a **diverging mirror**, which always produces diminished, upright and virtual images.

Key concept

The terminology, principles and equations involved with curved mirrors are very similar to those used for lenses.

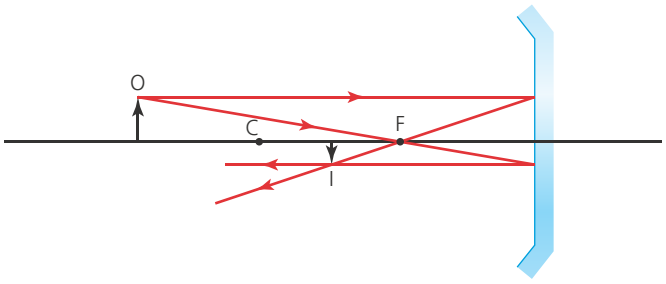


Figure 15.16

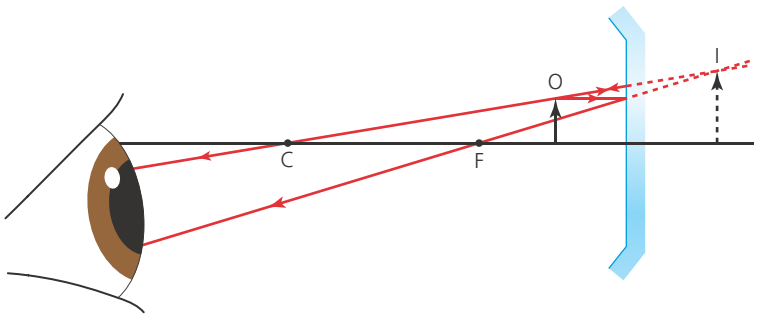


Figure 15.17

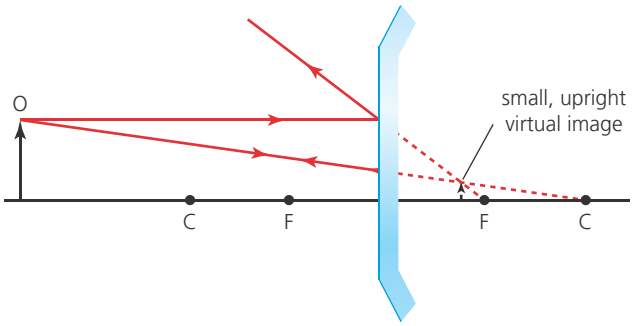


Figure 15.18

Solving problems involving not more than two curved mirrors by constructing scaled ray diagrams

- Figure 15.19 shows an example. (The principal axis of mirror A has been displaced to make the drawing clearer. The blue line is just a construction line used to locate the top of the final image.)

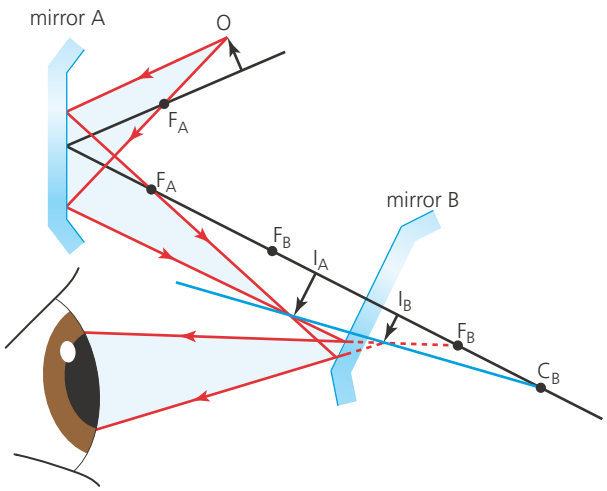


Figure 15.19

Spherical aberration in mirrors

- Figure 15.20 shows how a converging mirror with surfaces which are spherical is unable to form a point focus from parallel rays. This can be corrected by changing to a **parabolic reflector** (Figure 15.21).

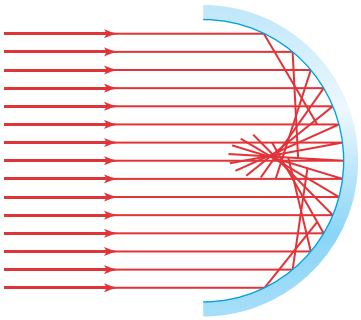


Figure 15.20

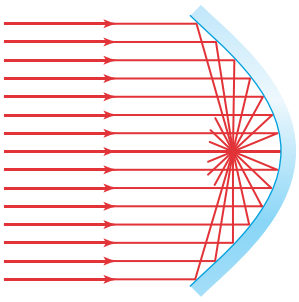


Figure 15.21

Expert tip

Rays from a point source placed at the focus of a parabolic surface will be reflected into a parallel beam, as shown in Figure 15.22. This is useful in spotlights, car headlights and anywhere else that a beam is to be directed from one place to another without spreading out, for example in a radar system.

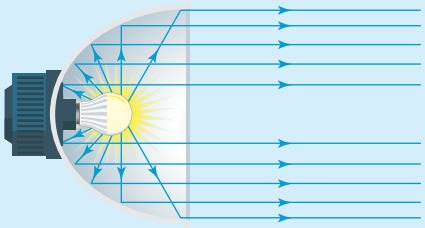


Figure 15.22

QUESTIONS TO CHECK UNDERSTANDING

17 a Draw a ray diagram to show the formation of an image by a converging mirror of focal length 12 cm when an object of height 2.0 cm is placed 16 cm in front of it.

b List the properties of the image.

18 Give one everyday example of a diverging mirror being used to produce diminished virtual images.

19 Explain why a parabolic converging mirror can produce better images than a mirror with spherical surfaces.

20 Make a copy of Figure 15.18, but make changes which will produce a larger final image with the same mirror and object.

NATURE OF SCIENCE

Deductive logic

The formation of real images by converging rays is readily understood from direct and shared observations, but the topic of *imaging* cannot be fully understood without extending that knowledge to the kind of images (virtual images) which are seen only by ourselves.

15.2 Imaging instrumentation

Revised

Essential idea: Optical microscopes and telescopes utilize similar physical properties to 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 Section 15.1 above, we saw how a single converging lens can produce magnified, virtual images of small objects placed close to it. For example a +10 D magnifying glass would have an angular magnification between 2.5 and 3.5.
- Trying to achieve greater magnification by using a more powerful lens will cause problems with the quality of the image because the greater curvature of the lens surfaces will make the effects of lens aberrations more noticeable.
- Greater magnification is possible by using lens combinations in *compound microscopes*.

Optical compound microscopes

Revised

- The simplest kind of optical microscope has two converging lenses.
- It is desirable for both lenses to be powerful, with small focal lengths (provided that lens aberrations are not significant).

Constructing and interpreting ray diagrams of optical compound microscopes at normal adjustment

- Figure 15.23 shows a ray diagram of a compound microscope. The object is positioned just beyond the focal point of the objective, so that it produces a real, magnified and inverted image. This first image is located closer to the eyepiece than its focal point, so that the eyepiece can act as a magnifying glass.
- In Figure 15.23, the microscope is shown at *normal adjustment* so that the final image is at the near point of the user's eye in order to achieve maximum magnification.
- When constructing this diagram, the top of the final image is located at the point where the 'construction line' through the centre of the lens meets the extension of the ray through the focal point.

Key concept

A simple **optical compound microscope** consists of two converging lenses. The lens close to the object (the **objective lens**) forms a real magnified image, and the second lens (the **eyepiece**) acts as a magnifying glass to further magnify the size of the final image, which is virtual and inverted.

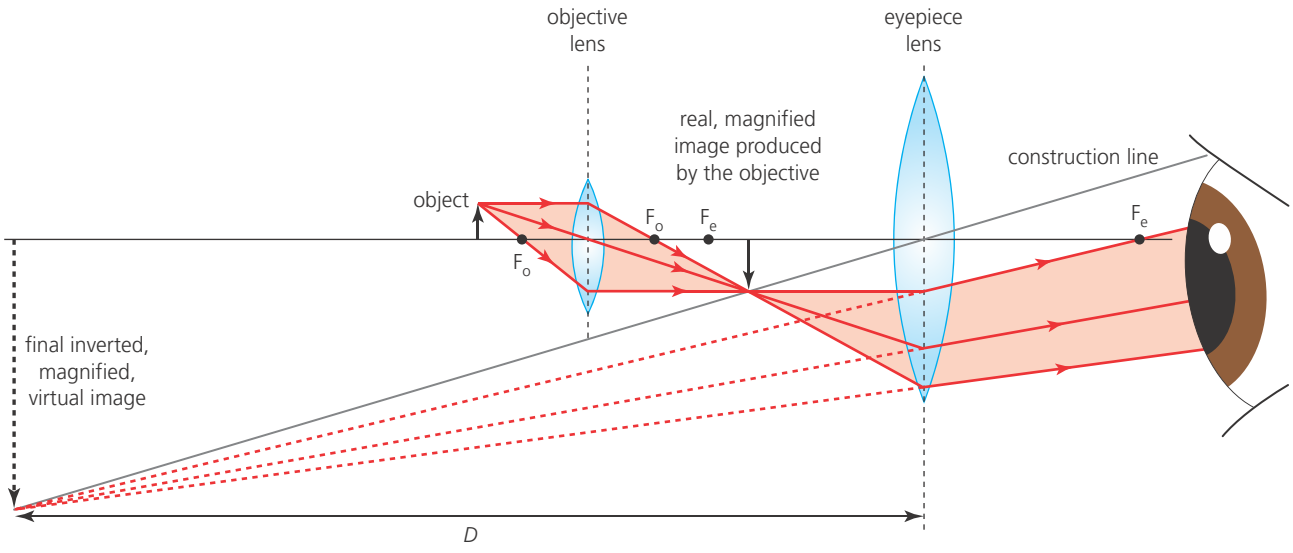


Figure 15.23

Investigating the optical compound microscope experimentally

- Figure 15.24 shows one way of investigating a model microscope in a darkened room. A translucent screen can be used to locate the image between the lenses.

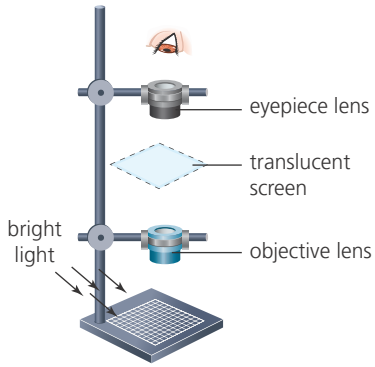


Figure 15.24

Solving problems involving the angular magnification and resolution of an optical compound microscope

- As an example, suppose an object was placed 3.5 cm in front of an objective lens of focal length 3.0 cm: the first image would be formed 21 cm from the objective lens and it would have a linear magnification $\left(m = -\frac{v}{u}\right)$ of −6.0. If the eyepiece lens had a focal length of 5.0 cm, it would need to be moved to a distance of 4.2 cm from the first image in order to create a final image at the near point. The angular magnification of the eyepiece lens $\left(M = \frac{D}{f} + 1\right)$ was 6.0, so that the overall magnification was 36.
- *Resolution* is usually more important than magnification in optical instruments. (Resolution was discussed for HL students in Chapter 9.)
- If the angle subtended at the eye by rays from two objects which can just be resolved is smaller, then the resolution is better (see Figure 15.25).

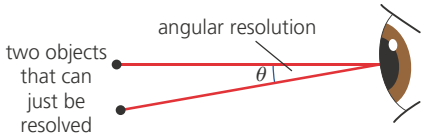


Figure 15.25

- Magnifying an image may improve resolution, but not if the resolution is already poor.
- In general, resolution is improved by using larger apertures and smaller wavelengths (these both minimize diffraction effects at the aperture).
- Better quality lenses will also significantly improve resolution.
- Large apertures (for example a large diameter objective lens) also have the advantage of collecting more light and producing brighter images.

Expert tips

An optical instrument may itself produce good resolution, but we may also have to consider the resolution of the eye or the camera which detects the image, or the resolution possible on the screen or photograph used to display an image.

Electron microscopes use the wave properties of electrons (Chapter 12 for HL students) instead of light. Because a typical electron wavelength is about 5000 times smaller than the wavelength of visible light, electron microscopes are capable of much greater resolution.

The resolution of a microscope can be improved by the use of a transparent oil between the specimen and the objective lens. This is because the oil has a higher refractive index than air.

Key concept

The angular magnification of a compound microscope equals the linear magnification of the objective lens multiplied by the angular magnification of the eyepiece lens.

Key concepts

An optical system may be described as producing good **resolution** if two close points on an object can be seen as separate on the image.

The actual resolution achieved can be measured as the angle subtended at the optical system by two points which can *just* be resolved. (A smaller angle means *better* resolution.)

The expected angular resolution of an optical system can be estimated from $\theta \approx \frac{\lambda}{b}$, where b is the diameter of the receiving aperture.

QUESTIONS TO CHECK UNDERSTANDING

21 a Draw a ray diagram to show the image formation by a compound microscope which has two lenses of focal lengths 3cm and 12 cm which are 30 cm apart.

b Estimate the angular magnification of this microscope at normal adjustment.

22 a The eyepiece of a compound microscope produces an angular magnification of 8.25 when the image is at the near point. What is the focal length of the eyepiece?

b If an object is placed 1.9 cm from the objective (focal length 1.5 cm) of this microscope, what is the overall magnification produced?

23 a Explain why you might expect a microscope with a larger diameter objective to produce better images.

b Suggest one problem that might occur when using larger lenses.

24 a Use the equation $\theta = \frac{1.22\lambda}{b}$ to estimate the minimum theoretical separation of two points that can be resolved on an object placed 1.5 cm from a microscope objective of diameter 1.0 cm. The factor 1.22 is included for circular apertures.

b Suggest why the actual resolution will be less than this value.

25 Discuss the advantages and disadvantages of using blue light, rather than white light, when observing a specimen through a microscope.

Optical telescopes

Revised

- We know from Section 15.1 that a single lens cannot be used to magnify a distant object.
- A *telescope* is an instrument for obtaining *angular* magnification of distant objects.
- Optical telescopes can be described as *refracting* or *reflecting*: a *refracting optical telescope* uses two or more converging lenses; a *reflecting optical telescope* uses one or more mirrors (and an eyepiece lens).
- Refracting telescopes with two converging lenses produce an *inverted* final image. Additional components are needed to produce upright images (if required).

Astronomical telescopes

- The use of visible light is just one part of astronomy. Light is only emitted by very hot astronomical sources. Most of the universe does not emit visible light.
- Waves from all parts of the electromagnetic spectrum arrive at the Earth from space. Different astronomical sources may emit different types of electromagnetic radiation.
- Accurately identifying the direction from which these waves have arrived is vital in 'mapping' the universe, and the waves also provide information about the nature of their sources. Radiation in different parts of the spectrum provide astronomers with different information.
- Figure 15.26 shows approximately how much radiation of different wavelengths reaches the Earth's surface.
- To maximize resolution, *terrestrial* astronomical telescopes should be placed as high as possible, which usually means on mountain tops. (**Terrestrial** means 'on the Earth').
- Note that there is a large part of the electromagnetic spectrum which is almost unaffected by passing through the atmosphere: radio waves and some microwaves. Terrestrial telescopes using these wavelengths do not have significant problems with absorption and scattering in the atmosphere. See section on Radio astronomy later.

Key concept

It should be clear from Figure 15.26 that significant amounts of most wavelengths are *absorbed* in the Earth's atmosphere. The atmosphere also *scatters* radiation and this can have a significant effect on the resolution of images from astronomical telescopes. Irregular refraction in the atmosphere can also cause problems for optical telescopes.

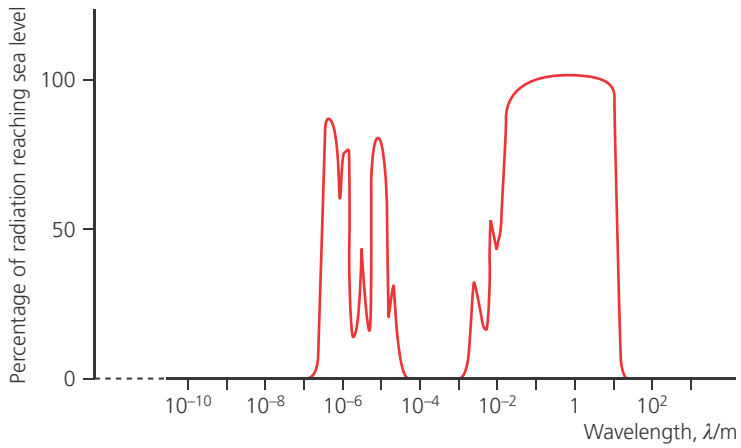


Figure 15.26

Satellite-borne telescopes

- Telescopes carried on satellites (*satellite-borne*) orbiting the Earth completely overcome the problems caused by the atmosphere for terrestrial telescopes.

Describing the comparative performance of Earth-based telescopes and satellite-borne telescopes

- It should be remembered that the latest developments in astronomical research often involve locating previously unrecorded, very distant sources. The amount of radiation received from such sources is tiny, so that it is very important that the telescopes are extra-sensitive and capable of high resolution.

Expert tips

The irregular refraction of light from a star passing through the Earth's atmosphere causes it to 'twinkle'.

Light from the Sun is scattered by particles as it passes through the Earth's atmosphere. This is why we can see the sky. A cloudless sky is blue because blue light is scattered more than other colours. On the Moon, which does not have an atmosphere, the sky is black.

Key concept

The main advantages of satellite-borne telescopes are (1) better resolution, (2) more sensitive, (3) unaffected by weather and pollution.

Simple optical astronomical refracting telescopes

Revised

- Figure 15.27 shows a ray diagram of a refracting telescope at **normal adjustment**, with the final image at infinity, allowing the eye to relax for prolonged observation.
- Note that the distance between the lenses is $f_o + f_e$.

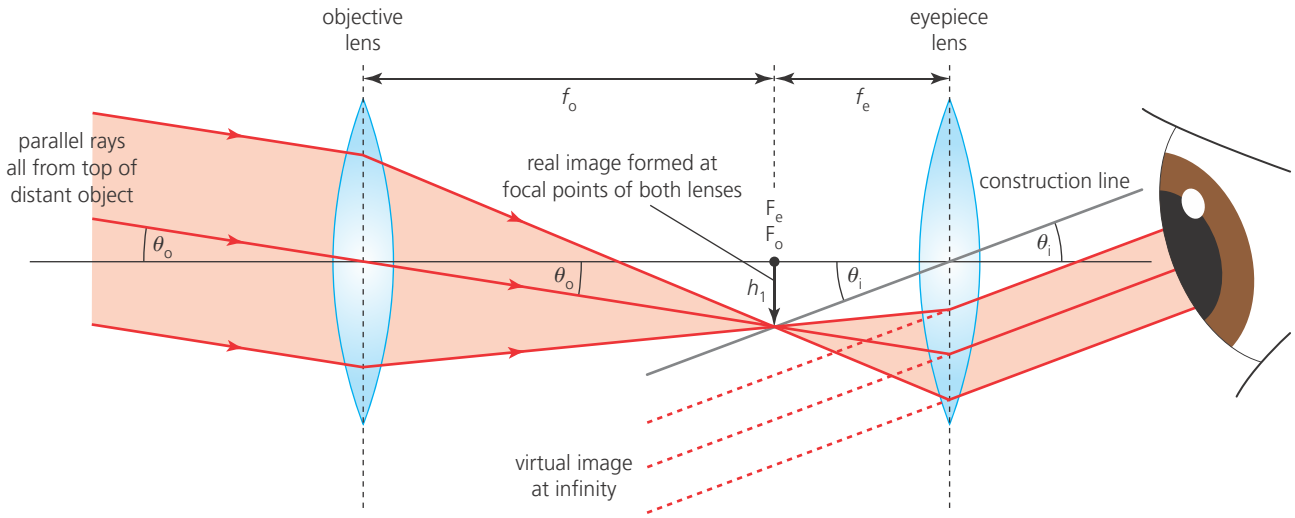


Figure 15.27

Constructing or completing ray diagrams of simple optical astronomical refracting telescopes at normal adjustment

- The 'construction line' through the top of the first image and the centre of the eyepiece lens is used to locate the direction of the top of the final image.
- From the geometry seen in Figure 15.27, it should be clear that the angular magnification, $M = \frac{\theta_i}{\theta_o} = \frac{f_o}{f_e}$.
- Greater magnification is achieved with an objective of longer focal length and an eyepiece of shorter focal length, but there are limitations on the power of the eyepiece before aberrations become significant.

Key concepts

The objective lens of a telescope forms a diminished, real and inverted image of a distant object at its focal point. The eyepiece then acts as a magnifying glass to produce a final image *at infinity* because the first image is also at the focal point of the eyepiece lens. The final image is inverted and virtual.

Angular magnification, $\mathbf{M = \frac{f_o}{f_e}}$.

Investigating the performance of a simple optical refracting astronomical telescope experimentally

Figure 15.28 shows one way of investigating a model refracting telescope in a darkened room. A translucent screen can be used to locate the image between the lenses.

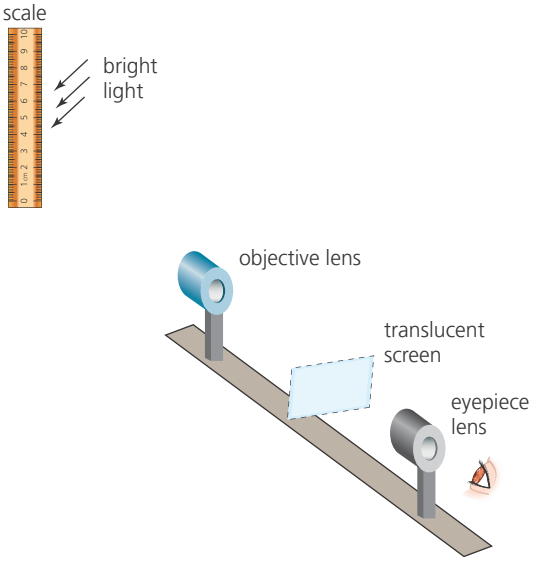


Figure 15.28

Simple optical astronomical reflecting telescopes

Revised

- Figure 15.29 shows an astronomical telescope which uses mirrors to collect and focus the light. This design, in which the observer looks into an eyepiece on the side of the telescope, is known as a **Newtonian mounting**.

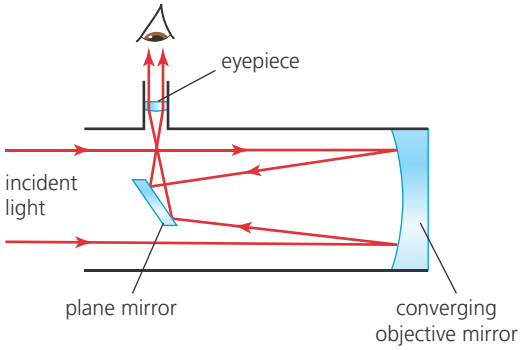


Figure 15.29

- The angular magnification of this telescope can be calculated from the same equation as for refracting telescopes.

- The disadvantages of this design are (1) that the plane mirror prevents some of the incident light reaching the converging mirror and (2) the observer is not looking towards the source.
- Figure 15.30 shows an alternative design, known as a **Cassegrain mounting**, in which the observer can look directly towards the source. The use of a diverging mirror also improves magnification.

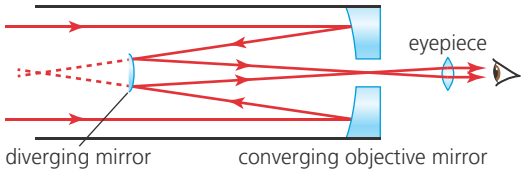


Figure 15.30

- Reflecting telescopes are considered to be better than refracting telescopes for most applications because light does not have to pass through any glass, in which there would be scattering and absorption, as well as chromatic aberration. A mirror can also be made with greater precision because it has only one important surface (a lens has two).
- Figure 15.31 shows the Hubble spacecraft in orbit around the Earth. It carried a 2.4 m diameter reflecting telescope.

Key concepts

Reflecting telescopes are generally considered to produce better images than refracting telescopes. Two common designs are called *Newtonian* and *Cassegrain*. They both use converging mirrors as objectives, but the Cassegrain mounting also involves a diverging mirror, which makes it easier to use.

Figure 15.31

Solving problems involving the angular magnification of simple optical astronomical telescopes

QUESTIONS TO CHECK UNDERSTANDING

26 List the advantages of placing optical telescopes on satellites.

27 a Draw a ray diagram for a refracting astronomical telescope at normal adjustment that has lenses with focal lengths of 5 cm and 20 cm.

b This would not be a very powerful telescope! What angular magnification would it produce?

28 What are the advantages and disadvantages of a telescope that has a large diameter objective?

29 What focal length of eyepiece will result in an angular magnification of 50 for a refracting astronomical telescope which has an objective lens of focal length 90 cm?

30 Consider a Newtonian reflecting telescope as shown in Figure 15.29. Give two reasons why this design is considered to be better than a telescope which uses two lenses to produce the same magnification.

Radio astronomy

Revised

- Optical astronomy has limitations. Light is only emitted by very hot astronomical sources, and it is affected by the atmosphere and the weather. Furthermore, optical astronomy is only possible at night (and even then it may be affected by light pollution).
- Unlike the other parts of the spectrum, the arrival of radio waves from space at the Earth's surface is mostly unaffected by the atmosphere or its weather, or the difference between night and day. Radio waves are also very useful in astronomy because they are detected from cool sources that do not emit visible light. They can also pass though intergalactic dust clouds without being affected. For example, radio waves of wavelength 21 cm are received from hydrogen throughout the universe.

Single dish radio telescopes

- Figure 15.32 shows the Parkes radio telescope in New South Wales, Australia.
- Because radio waves from space might have a typical wavelength of about 10 cm, good resolution using a single dish can only be achieved if it has a large diameter.
- There is a limit to the size and resolution of single dish radio telescopes because of the problems of precisely maintaining their shapes.

Figure 15.32

Key concept

The simplest radio telescopes have an **aerial (antenna)** placed at the focal point of a single parabolic dish reflector.

As with other instruments, the angular resolution of a radio telescope dish of width b is limited to angles larger than $\theta \approx \frac{\lambda}{b}$

Resolution can be improved by combining the signals received by an *array* of smaller telescopes.

Radio interferometry telescopes

- Higher resolution when receiving radio waves is possible using *interferometry techniques*, in which the signals from two or more synchronized telescopes are combined electronically.
- The signals *interfere* and the spacing and centre of the interference pattern can be used to accurately determine the direction to the source of radiation.
- However, by using many telescopes in an *array*, the resolution can be greatly increased. A *telescope array* is a regular arrangement of telescopes in one or more rows, as in Figure 15.33.

Key concept

Generally, the resolution of an array of smaller telescopes can be considered to be equivalent to a single large dish with diameter equal to the maximum separation of the telescopes (the 'baseline').

Figure 15.33

QUESTIONS TO CHECK UNDERSTANDING

31 Why can radio telescopes be used during the day, when optical astronomical telescopes cannot?

32 a What resolution can be achieved by the Parkes radio telescope when receiving waves of frequency 1666 MHz (diameter of dish is 64 m)?

b Compare your answer to the resolution of an optical telescope with an objective diameter of 2.7 cm.

33 The Very Large Array is a radio interferometer observatory in New Mexico, USA. There are 27 dishes each of diameter 25 m arranged in a Y shape. It is said to have a sensitivity equal to a single dish aerial of diameter 130 m.

a Check this claim by comparing the receiving areas of the 27 telescopes to a single dish of diameter 130 m.

b Estimate a resolution that might be achieved with this array.

34 The new Chinese FAST telescope (see Figure 15.34) has a single dish with a diameter of 500 m. Previously, the Arecibo radio telescope was the world's largest (diameter 305 m). Compare the resolution and sensitivity of these telescopes.

Figure 15.34

NATURE OF SCIENCE

Improved instrumentation

Improvements in the designs of microscopes and telescopes (including ways of capturing and processing images) have led to enormous advances in recent years in our knowledge of both very small and the very distant objects.

15.3 Fibre optics

Revised

Essential idea: 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 transmission of data along cables

Revised

- For personal communications, the use of radio waves (microwaves) through the air is convenient and commonplace, but *cables* are used for most of the data which is transmitted around the world.
- Most data is transmitted using either electrical pulses in copper cables, or infrared pulses in **optic fibres**.
- Data is usually sent using **digital pulses**, rather than continuously varying *analogue signals*. Digital data is transferred as a very large number of individual pulses, each of which can have only one of two possible levels. (The levels in such a *binary* system are commonly called 0 or 1). Figure 15.35 shows the intensity of a series of digital pulses that might be sent along a cable. (They represent the **binary number** 11010010, which is equal to the decimal number 210.)

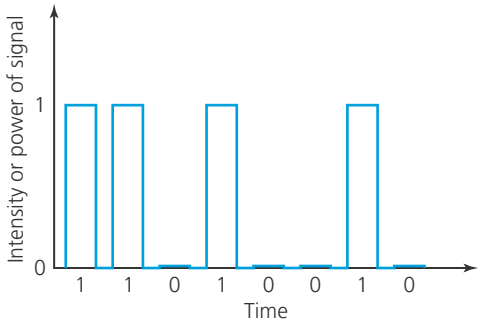


Figure 15.35

- As pulses travel along a cable they *attenuate* and *disperse*.
- Figure 15.36 illustrates the problem caused by dispersion: pulses which were clearly separate may overlap after they have travelled some distance along a cable.

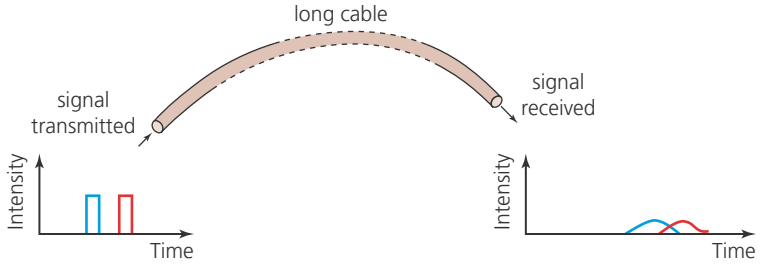


Figure 15.36

- Data is transferred in digital form because when pulses are affected by the problems of dispersion, attenuation and 'interference' (see below), they can still usually be distinguished as 0 s or 1 s because there are only two very different levels (whereas small changes in the amplitude of analogue signals can be misinterpreted).
- Attenuation and dispersion limit the distance that data can be transferred (before it needs to be amplified and reformed) and the amount of data that can be sent in a given time through a particular cable.

Expert tip

The terms *wire* and *cable* are commonly confused. A copper *wire* is a single conductor which may or may not be insulated, whereas a copper *cable* consists of two or more conductors which are well insulated from each other. In both cases, the conductor(s) may be a single strand or a group of strands twisted together (to make them more flexible). An optic fibre in an optic cable has protective layers around it and there may be many fibres in the same cable.

Key concepts

Attenuation is the gradual loss of intensity of a signal as it passes through a material.

Dispersion is the broadening of the duration (width) of a pulse (and its consequential attenuation).

Data transmission in copper cables

- This topic concentrates on the advantages of optic fibres for data transmission, but we will begin with a review of the use of *electrical pulses in copper wires*. Data is transmitted along a cable as pulses of varying potential difference between two wires. There will be associated electromagnetic waves between and around the wires.
- These changing electromagnetic fields can spread away from the cable and cause **electronic 'interference'** by inducing tiny emfs in other cables (Chapter 11). Such effects may be described as an example of unwanted electronic '*noise*'. (These effects should be distinguished from the *interference* of waves.)
- Electronic interference can be significantly reduced in copper cabling by twisting wires together, which are then known as *twisted pairs* (Figure 15.37 shows four twisted pairs in a cable), or by using *co-axial cables*, as shown in Figure 15.38. The outer copper mesh is connected to earth (0 V) and this limits the passage of electromagnetic waves through it to the central conductor.

Key concepts

Twisted pairs (of wires) and **coaxial cables** are used to reduce the amount of electrical interference between wires carrying electrical signals.

Figure 15.37

Figure 15.38

Describing the advantages of fibre optics over twisted pair and coaxial cables

- The use of fibre optic cables to transfer data has many obvious advantages over copper wiring.
- (The installation costs of optic fibres, including the necessary connections to copper wiring, are relatively high.)

Key concept

The effects of dispersion, attenuation and interference are much less significant when optic fibres are used to transmit data. Optic fibres are also able to transfer more data in the same time (greater *bandwidth*) than copper wires of similar dimensions. They are smaller and lighter in weight, and considered to be more private/secure.

QUESTIONS TO CHECK UNDERSTANDING

35 Explain why when a pulse is dispersed, its intensity must also decrease.

36 Give another example of wave attenuation (apart from that occurring in optic fibres).

37 (Higher level) Use knowledge from Chapter 11 to explain how a changing current in one wire can induce a voltage in another wire.

Data transmission in optic fibres

Revised

- Optical data transmission relies on *total internal reflection*. This effect, shown in Figure 15.39, was discussed in Chapter 4, Section 4.4, and is briefly revised below.
- The radiation is reflected internally when the angle of incidence inside the fibre is larger than the critical angle.

Key concept

The transmission of data using pulses of infrared radiation along optic fibres (made from high purity glass) is only possible because of the phenomenon of total internal reflection.

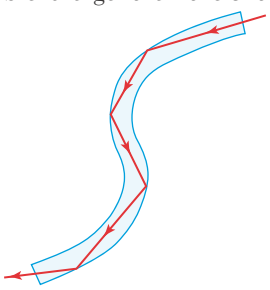


Figure 15.39

Total internal reflection and critical angle

Revised

- When a ray passes into another transparent medium in which it would travel slower (greater refractive index, n) it can refract away from the normal. At a particular angle, called the *critical angle*, c, the angle of refraction is 90° as shown in Figure 15.40.

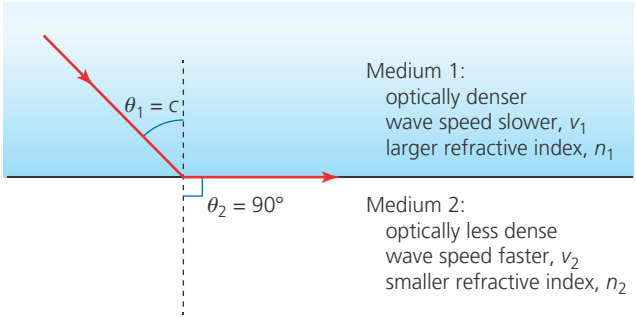


Figure 15.40

- From the IB 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 into air (medium 2) then $n_2 = n_{air} = 1$, so that (replacing n_1 with n): $n = \frac{1}{\sin c}$.

Expert tip

Ideally, a light ray entering an optic fibre needs to be parallel, or close to parallel, to the axis of the fibre to be sure that it will be totally internally reflected, but larger angles are acceptable. The greatest possible angle of incidence to the axis is called the *acceptance angle*.

Structure of optic fibres

Revised

- The typical structure of a single core optic fibre(s) is shown in Figure 15.41.
- The inner glass fibre is surrounded by glass of a lower refractive index. This is called *cladding*; it has three purposes:
  - Because it has a refractive index which is greater than air, the critical angle when cladding is used is greater, which means that only light rays striking it with relatively large angles of incidence will be reflected. Only those rays that are travelling close to the axis of the fibre pass can be transmitted. (See discussion of dispersion below.)
  - The inner fibre is protected from damage.
  - If the cable is multi-cored, the inner fibres cannot come in contact with each other. This prevents light passing between fibres ('crosstalk').
- The outer layers add strength and protection against the environment.

Key concept

The inner glass fibre in an optic cable is surrounded by another layer of a different glass, called **cladding**. The cladding protects the core and prevents different cores from coming in contact with each other.

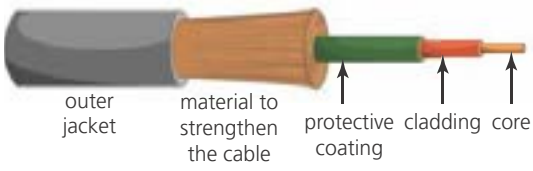


Figure 15.41

Solving problems involving total internal reflection and critical angle in the context of fibre optics

QUESTIONS TO CHECK UNDERSTANDING

38 a What is the critical angle for light travelling in an optic fibre made from glass of refractive index 1.55 when it is surrounded by:

i air

ii another type of glass of refractive index 1.48?

b Explain the difference that the glass cladding makes to the transmission of light along the fibre.

39 Sketch rays entering an optic fibre to illustrate the concept of *acceptance angle*.

Waveguide and material dispersion in optic fibres

Revised

Refer back to Figure 15.36. Dispersion in optic fibres has two main causes – *material dispersion* (sometimes called 'modal dispersion') and *waveguide dispersion*.

Describing how waveguide and material dispersion can lead to attenuation and how this can be accounted for

- *Material dispersion* will occur if radiation of different wavelengths is used. This is because they travel at very slightly different speeds (so they have slightly different refractive indices). This can be overcome by using infrared with a very narrow range of wavelengths (from an LED).
- Data is transferred in optic fibres using *infrared radiation* because most of these wavelengths have low absorption and scattering in glass.
- *Waveguide dispersion* is due to the fact that different rays (that started together) travel along slightly different paths as they are transmitted along the fibre. This is represented in Figure 15.42 (but the effect has been exaggerated for clarity).

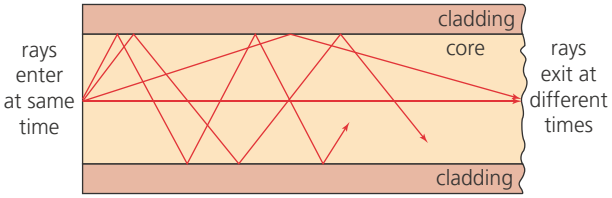


Figure 15.42

Key concepts

There are two main causes of dispersion in an optic fibre. They both result in the possibility of waves taking different times to travel the same distance along a fibre.

1 **Material dispersion**: if different wavelengths are used they will travel at slightly different speeds. This problem can be overcome by using *monochromatic radiation*.

2 **Waveguide dispersion**: waves transmitted at slightly different angles will follow paths of slightly different length. This problem can be reduced by using *cladding* and *graded index fibres*.

Expert tip

The term **waveguide** is used to describe any structure which has been constructed in order to transmit waves along a particular path with minimal loss of intensity. Apart from optic fibres, the term is most commonly used for microwaves, radio waves and sound.

Step-index fibres and graded-index fibres

- Waveguide dispersion can be reduced by using *graded-index fibres* as shown in Figure 15.43.
- *Graded-index fibres* should be compared to **step-index fibres**, which have fibres of constant refractive index. Figure 15.44 shows the difference between them.

Key concept

In **graded-index fibres**, 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, so that they all take similar times to travel the same length of fibre, thus reducing waveguide *dispersion*.

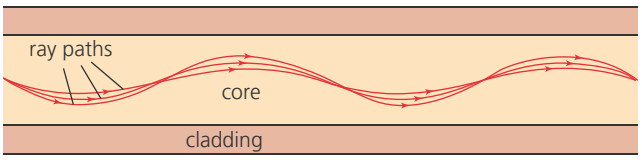


Figure 15.43

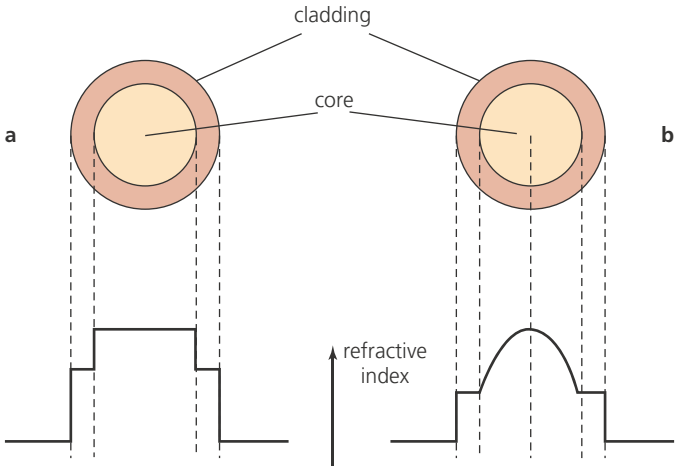


Figure 15.44 Cross-sectional variation of refractive indices in (a) step-index fibres and (b) graded-index fibres.

QUESTIONS TO CHECK UNDERSTANDING

40 Some optic fibre cable cores may have a diameter smaller than 1.0×10^{-5} m.

- **a** Approximately how many wavelengths (of the radiation used to transmit data) is this?
- **b** Suggest a reason why it may be better for an optic fibre to have a very small diameter.
- **c** Explain why a *ray model* for the transmission of the radiation (as used in the figures in this section) may not be appropriate for very thin fibres.

41 The refractive index for infrared radiation of wavelength 1.5×10^{-6} m in a certain type of glass is 1.5578, while for a wavelength of 2.0×10^{-6} m the refractive index in the same glass is 1.5516. Determine how much longer it takes for the shorter wavelength to travel 100.0 m in the glass (speed of light in vacuum = $2.9979 \times 10^{8}\,\text{m s}^{-1}$.

Attenuation and the decibel (dB) scale

Revised

- As stated above, *attenuation* is the gradual loss of intensity of a signal as it passes through a material. Attenuation in an optic fibre is caused by *dispersion* (as the duration of a pulse increases, its intensity decreases) and also by **scattering** from very small irregularities in the glass and **absorption** due to impurities.
- Attenuation is frequency dependent, as shown by the example for a certain type of glass in Figure 15.45.

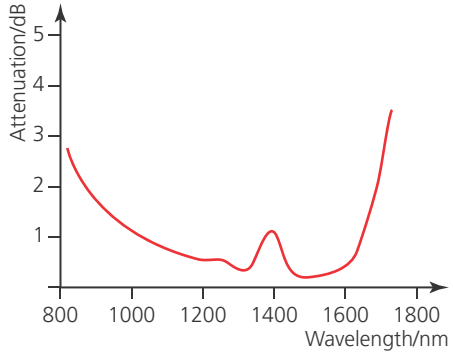


Figure 15.45

Key concepts

Attenuation in an optic fibre is due to absorption, scattering and dispersion.

Intensity decreases exponentially with distance along the fibre, and the attenuation can be calculated from: $\text{attenuation} = 10 \log \left(\frac{I}{I_0} \right)$.

The unit for attenuation calculated in this way is the decibel, dB.

Attenuation is commonly quoted in dB per unit length.

- As a signal travels along an optic fibre, its intensity decreases by equal percentages in equal distances. For example an intensity may fall to 80% (or by 20%) every 10 km. This means that the decrease in intensity is *exponential* and requires the use of logarithms in equations to fully represent it. (Similar to the mathematics of radioactive decay and the discharge of capacitors.)
- If the intensity decreases from I_0 to I between two points on the fibre, then the attenuation can be calculated from $\log\left(\frac{I}{I_0}\right)$. For example, if the intensity fell by 70%, the attenuation would be –0.52. Note that attenuation values are always negative because I is always less than I_0. The unit of attenuation calculated in this way is the *bel*, B.
- However, it is considered more convenient to use larger numbers and the usual units for attenuation are **decibels** (dB), then attenuation (dB) = $10\log\left(\frac{I}{I_0}\right)$.

  The IB Physics course does not use any symbol to represent attenuation. The attenuation in the previous example (−0.52 B) is −5.2 dB.
- It is usual to give a value for the attenuation *per unit length* of a cable. As such, it may be called the *attenuation coefficient*. For example, if the attenuation coefficient in an optic fibre was quoted to be –0.32 dB km^{-1}, the intensity of a signal in that fibre would decrease by 7.1% in a distance of 1 km. Note that, since the decrease is exponential, this should *not* suggest that the decrease in 2 km would be 14.2% (it would be 13.7%).
- Exponential decreases in *power* can be described by a similar equation.
- The attenuation in optic fibres is typically –1 dB km^{-1} (to an order of magnitude), compared to –100 dB km^{-1} for twisted pairs or coaxial cable.
- Refer back to Figure 15.36. If it is required to send signals over large distances, it will be necessary to restore the intensity and shape of the pulses at suitable distances along the cable. This is called *regeneration*.

Solving problems involving attenuation

QUESTIONS TO CHECK UNDERSTANDING

42 Use Figure 15.45 to choose two suitable wavelengths for data transmission in glass.

43 What is the attenuation (dB) in a length of optic fibre if the intensity falls by 90% between its ends?

44 If the overall attenuation in a cable is –0.5 dB, by what percentage is the output intensity decreased compared to the input?

45 An optic fibre cable is rated at –1.25 dB km^{-1}.

- **a** Determine the power after a length of 1.0 km if the power input was 20 mW.
- **b** Calculate the power loss after a total distance of 2.0 km.
- **c** Express the attenuation in the cable in dB after a distance of 5.0 km.

46 If a signal needs to be regenerated after its intensity has fallen to 20%, estimate the maximum distance between regenerators (repeaters) in a cable which has an attenuation of –1.5 dB km^{-1}.

Expert tip

The decibel (dB) scale is commonly used to *compare* an intensity (or power) to a well-defined reference level. For example, a very quiet sound may be described as having an intensity of 10 dB, which means, by definition in this example, that the intensity is 10 times more intense than the agreed intensity of the threshold of human hearing, which is said to be 0 dB. A sound of 20 dB is 100 times more intense than 0 dB, a sound of 30 dB is 1000 times more intense than 0 dB, etc. A loud sound of 100 dB is 10^{10} times more intense than the threshold of hearing.

Expert tip

Thin flexible optic fibres are also widely used for obtaining images of inaccessible places, such as inside engines. A bundle of fibres has an objective lens at one end and an eyepiece, or camera, at the other. Typically, light is needed to illuminate the object being examined and this is carried down some of the fibres. Figure 15.46 shows a medical use of such a device. It is called an endoscope.

Figure 15.46 An endoscope can be used to inspect a patient's stomach.

NATURE OF SCIENCE

Applied science

We have become dependent on very fast worldwide digital communication and this has been accomplished largely because of the use of fibre optics (much of which is unseen by the general public). This has mostly been due to the application of new and improved technologies to ideas about total internal reflection that have been well understood for hundreds of years.

15.4 Medical imaging (Additional higher level)

Revised

Essential idea: 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.

X-rays

Revised

- X-rays are electromagnetic radiation within the approximate range of wavelengths 1×10^{-11} m to 1×10^{-9} m. When we refer to X-rays we are usually talking about artificially produced radiation, but X-rays also have some natural origins. The production of X-rays in 'X-ray tubes' is not part of the IB Physics course.
- X-rays and gamma rays of the same wavelength are identical to each other.
- As an example, an X-ray of wavelength 2.0×10^{-11} m has a frequency, $f = \frac{c}{\lambda} = \frac{(3.0 \times 10^{8})}{(2.0 \times 10^{-11})} = 1.5 \times 10^{19}$ Hz. The energy of this photon, $E = hf = 6.63 \times 10^{-34} \times 1.5 \times 10^{19} = 9.9 \times 10^{-15}$ J. However, photon energies are more usually quoted in eV, so that $E = \frac{(9.9 \times 10^{-15})}{(1.6 \times 10^{-19})} = 6.2 \times 10^{4}$ eV, or 62 keV.
- X-rays with shorter wavelengths (and more energy) are more penetrating than longer wavelengths, and they are sometimes described as 'hard' X-rays. Typical 'hard' X-ray energies used in diagnostic investigations in hospitals are 20–100 keV.
- An X-ray beam will typically contain a continuous range of different wavelengths.

Expert tip

From Section 12.1, we know that gamma ray (or X-ray) photons of energy greater than 1022 keV can cause *pair production*, but these energies are much higher than used for medical X-rays.

QUESTIONS TO CHECK UNDERSTANDING

47 What is the wavelength of a soft X-ray which has an energy of 10 keV?

48 Explain the difference between X-rays and gamma rays.

Explaining features of X-ray imaging

Revised

- The use of X-rays to obtain images of the internal structures of bodies is a major function of the *radiography* departments of hospitals.
- 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.

Key concept

Different parts of the body will absorb X-rays by different amounts and the intensity of the transmitted beam will show variations representing the presence of parts of the body with different densities and absorption rates. An example is shown in Figure 15.47.

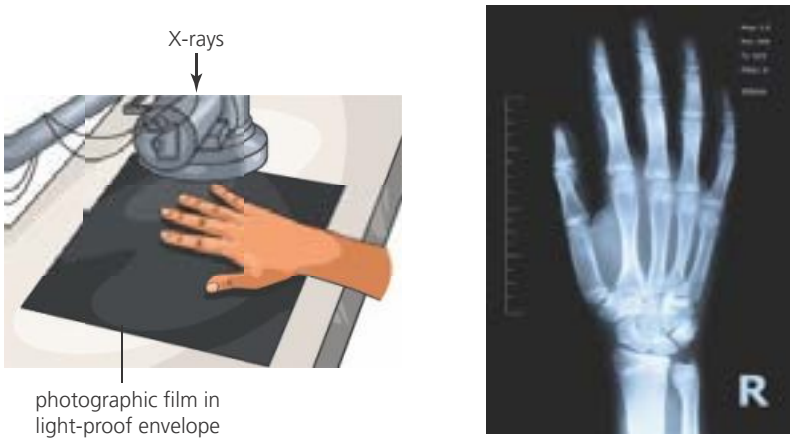


Figure 15.47

Detection and recording of X-ray images in medical contexts

- The X-rays that are transmitted can be detected either photographically, as shown in Figure 15.47, or (preferably) electronically, for example, by the use of **charge-coupled devices** (CCDs) as used in digital cameras.

Key concept

The use of electronic detectors facilitates the storage and manipulation of images. The images have better resolution than photographic images and they are available immediately (requiring no further processing). Furthermore, because of the greater sensitivity of the detectors, electronic procedures involve lower power and safer X-ray beams. The power of the beam may be controlled automatically to the minimum required.

Attenuation of X-rays

Revised

- Reminder: *attenuation* is the gradual decrease in intensity as a beam passes through a medium.
- X-ray absorption is an ionizing processes which can result in chemical and biological changes which will be harmful to the human body. Doctors need to balance the (small) possible risk to the patient of using X-rays against the benefits of correctly diagnosing a medical problem. Hospital staff also need to be protected.

- Attenuation decreases significantly with increasing frequency (energy) of the X-rays being used. As stated before, more energetic X-ray photons are more penetrating.
- Absorption due to the photoelectric effect increases significantly with the proton numbers, Z, of the elements present. This explains why, for example, bone absorbs more than soft tissue.
- Figure 15.48 shows a simplified representation of the variation of *mass attenuation coefficient* (explained below) with photon energy for two different materials. Note that the variation is greater than it may at first appear because the vertical scale is logarithmic.

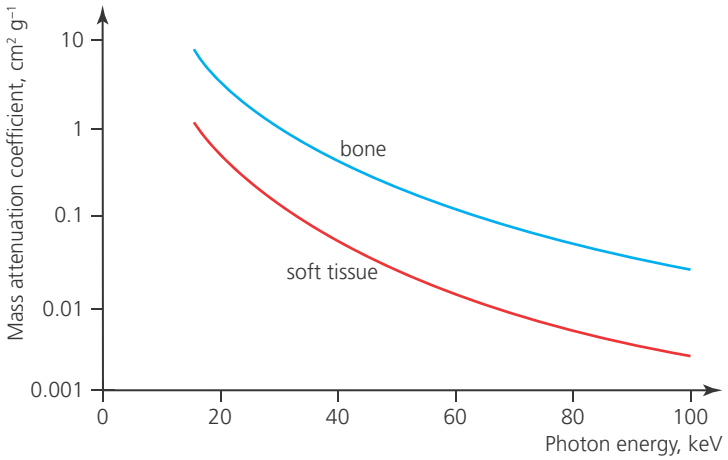


Figure 15.48

Key concepts

X-rays are mainly attenuated because of scattering and because of absorption due to the *photo-electric effect*, which is greater in materials of higher proton number.

More energetic X-rays are more penetrating: attenuation is less for X-rays of greater frequencies ('*harder*' X-rays).

Representing the attenuation of X-rays mathematically

- The equations below make these simplifying assumptions:
  - The X-ray beam has photons of only one wavelength (energy).
  - The beam is not spreading out; it is parallel.
  - The beam is passing perpendicularly through a single homogeneous medium.
- Figure 15.49 shows a beam of intensity I_0 which has its intensity reduced to I after passing through a material of thickness x.
- In practice, an X-ray beam may have a range of energies. The less energetic photons will be more easily attenuated, resulting in the beam becoming 'harder' as its penetration increases.
- Attenuation can be represented in the same way as for attenuation in an optic fibre: $\text{attenuation (dB)} = 10 \log\left(\frac{I}{I_0}\right)$.

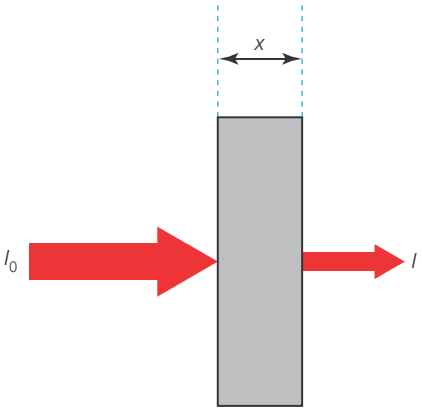


Figure 15.49

QUESTIONS TO CHECK UNDERSTANDING

49 Explain why a greater percentage of X-rays will pass through soft tissue than an equal thickness of bone.

50 Use Figure 15.48 to determine by what factor the absorption of 40 keV X-rays is greater in bone than soft tissue.

51 When X-rays pass through some muscle their intensity is reduced by 65%. Calculate the attenuation in dB.

Explaining attenuation coefficient, half-value thickness and linear/mass absorption

- The intensity of a parallel beam of X-rays decreases exponentially with distance, x, due to absorption and scattering. This is shown by either of the curves in Figure 15.50.
- This exponential decrease is represented by the equation $I = I_o e^{-\mu x}$. Where μ is a constant called the **linear attenuation coefficient** for radiation of a specified wavelength (energy). Usual unit: cm^{-1}.
- For example, the linear absorption coefficient for X-rays of energy 60 keV in bone is about $0.50\,cm^{-1}$. This means that the intensity of this particular X-ray wavelength after passing through, for example, a thickness of 1.0 cm of bone can be calculated from $I = I_o e^{-(0.50 \times 1.0)}$, which corresponds to a decrease to 61% of the incident intensity.
- Figure 15.50 illustrates graphically the effect of linear attenuation coefficient on absorption. Figure 15.51 illustrates the exponential variation of intensity with thickness in more detail.
- The attenuation of X-rays is also often characterized by the **half-value thickness** of a particular medium, $x_{\frac{1}{2}}$. This concept is similar to the use of *half-life* in radioactivity. Its meaning is shown graphically in Figure 15.51.

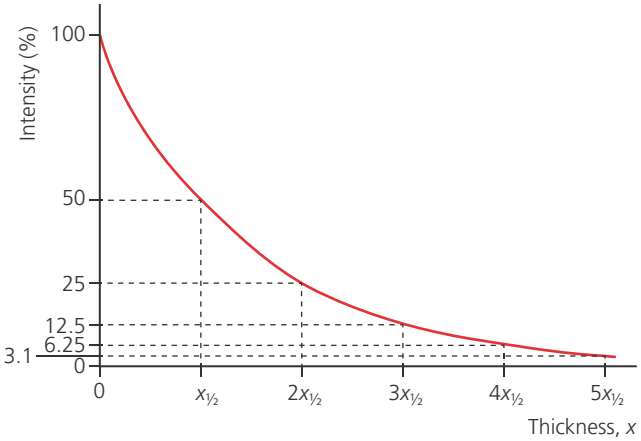


Figure 15.51

- Since $\frac{I}{I_o} = e^{-\mu x}$, it should be clear that $0.5 = e^{-\mu x_{\frac{1}{2}}}$, so that $\mu x_{\frac{1}{2}} = \ln 2$.
- The *mass attenuation coefficient* is commonly used to compare the attenuation in unit masses of different materials. **Mass attenuation coefficient** $= \frac{\text{linear attenuation coefficient}}{\text{density}} = \frac{\mu}{\rho}$.

Solving X-ray attenuation problems

The equations highlighted above can be used to determine the intensity of X-rays transmitted through a known thickness of material if the linear attenuation coefficient is known (or the mass attenuation coefficient and the density of the material). Calculations will assume that the surfaces are parallel. If two or more layers of different materials are involved, they must each be treated separately.

QUESTIONS TO CHECK UNDERSTANDING

52 a Why does the intensity of a parallel beam of X-rays decrease exponentially with distance travelled through a medium?

b Muscle density is $1.06 \times 10^3\,kg\,m^{-3}$. If its mass attenuation coefficient is $0.227\,cm^2\,g^{-1}$, calculate the linear attenuation coefficient of muscle tissue.

c Your answer to (b) applies to only to a particular X-ray energy. How would your answer change if X-rays of smaller wavelength were used?

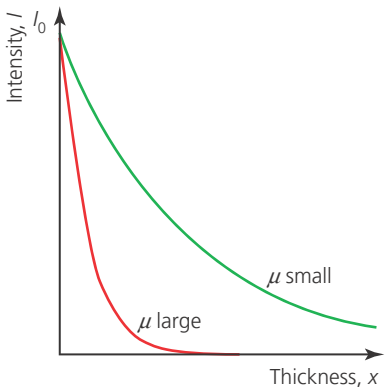


Figure 15.50

Key concepts

The linear attenuation coefficient, μ, of a material represents the amount of attenuation of X-rays in unit thickness (usually per centimetre).

$I = I_o e^{-\mu x}$

Alternatively, attenuation may be represented by the thickness of a material which halves the intensity: the half-value thickness,

$x_{\frac{1}{2}}$, $\mu x_{\frac{1}{2}} = \ln 2$

Mass attenuation coefficient $\left(= \frac{\mu}{\rho}\right)$ is also used to compare attenuation in different materials. Usual units: $cm^2\,g^{-1}$.

53 If X-rays of intensity I_0 are incident upon a certain bone of thickness 14 mm, what is the intensity of the radiation which emerges if the linear absorption coefficient is 0.52 cm^{-1}?

54 The half value thickness for a particular medium (at a given wavelength) is 1.18 cm.

a What is the linear attenuation coefficient of this medium?

b The same medium has a mass attenuation coefficient of 0.534 $cm^2 g^{-1}$. What is its density?

c Determine what thickness of this medium will reduce the intensity of an X-ray beam to 10%.

55 Blood has a density of 1.02 g cm^{-3} and a mass attenuation coefficient of 0.18 $cm^2 g^{-1}$ (for 80 keV X-rays). Estimate by how much the X-ray intensity is reduced by passing through a blood vessel of diameter 2.2 mm.

56 Figure 15.52 shows a parallel beam of X-rays passing through the tissue and bone (femur) of a thigh (not to scale). Use the data in the figure to calculate by what percentage the incident intensity has fallen by the time it emerges at A and B.

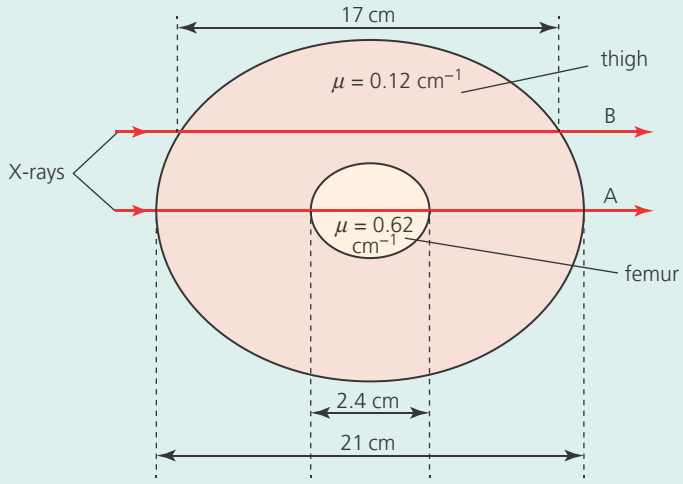


Figure 15.52

Explaining techniques for improvements of sharpness and contrast

Revised

- It is obvious that the images obtained with X-rays should show as much detail as possible (*high resolution*). The small wavelength of X-rays means that diffraction is not a problem.
- High-quality X-ray images should also be *sharp* and have good *contrast*.
- The X-rays are not focused to form an image, so it is important for sharp images with good contrast that (1) the X-ray sensors on the detector are close together and (2) that all the radiation reaching any particular place on the detector has followed the same single straight line path from the X-ray source and through the patient, that is, the X-ray source is small and scattered radiation does not reach the detector. Figure 15.53 shows the direct path of one ray.

Key concepts

When we describe an image as **sharp** we mean that any edges between different parts of the image are distinct and precise.

Different parts of an image with good **contrast** have significantly different brightnesses (or colour).

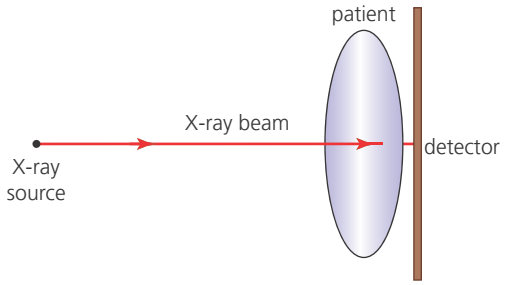


Figure 15.53

- The sharpness of images improves if the source–patient distance is larger and/or the patient–detector distance is smaller, and if there is no movement of the patient.
- Lower energy X-rays are more easily scattered in a patient, so an aluminium filter is often placed in the X-ray beam to remove most of the longer wavelengths.
- Figure 15.54 shows other techniques. The X-rays pass through an absorbing grid with holes. Each hole acts as a **collimator** and only allows parallel X-rays to pass through it. The grid may have to oscillate to allow coverage of all of the patient. **Intensifying fluorescent screens** convert X-rays into visible light, which is more easily detected.
- For some diagnoses, a 'contrast medium', which has a high absorption coefficient, is temporarily introduced into the body to improve contrast.
- The quality of images may also be enhanced by computer programmes, especially in CT scans (see below).

Key concepts

Various techniques can be used to improve sharpness and contrast in X-ray images. These include (1) increasing source–patient distance; (2) using a filter to absorb low energy X-rays; (3) using a collimating grid; (4) using a fluorescent screen intensifier; (5) enhancing images with a computer programme.

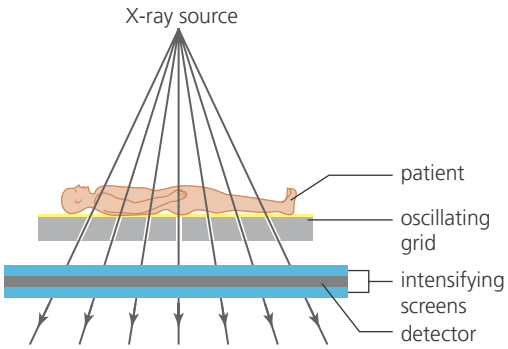


Figure 15.54

QUESTIONS TO CHECK UNDERSTANDING

57 Make a list of the various ways in which X-ray images can be made to have better resolution, sharpness or contrast.

58 Explain, with the help of a diagram, why a larger X-ray source will produce less sharp images.

59 Explain why the small wavelength of X-rays helps produce images with good resolution.

Computed tomography

Revised

- Tomography is a technique for displaying a cross-section through a human body or other solid object using X-rays which are directed at the patient from different angles.
- Computed tomography (CT) scans can be combined to present a three-dimensional image. 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%. CT scans involve much longer times and the patient has a greater exposure to radiation than with conventional X-rays.

Key concept

Computed tomography (CT) uses computer-controlled X-rays and rotating machinery to obtain sharp images of planes of the patient (scans).

Medical ultrasound

Revised

- Figure 15.55 shows a patient having an abdominal ultrasound scan.
- Ultrasound imaging has no known risk, but the images can have disappointing resolution because of the relatively long wavelengths used. Ultrasound cannot penetrate into bone effectively and cannot be used for spaces that contain air (e.g. lungs).

Figure 15.55

Ultrasound waves

Revised

- **Ultrasound** is longitudinal sound waves with frequencies higher than can be heard by humans.
- Like all waves, ultrasound will also be *refracted* and *diffracted* under suitable circumstances.

Explaining features of medical ultrasound techniques

- A typical frequency of ultrasound in medical use is 2 MHz, which corresponds to a wavelength of about 1 mm at a speed of 1540 $m\,s^{-1}$, which is a typical speed for inside the human body.
- The waves are generated by an ultrasound **transducer** (sometimes called a **probe**) and directed into the body. Some of the waves are reflected off various boundaries and arrive back at the same transducer (like an *echo*), which also acts as a detector. A *transducer* is the general term for any device that converts variations in a physical quantity, such as force or intensity, into an electrical signal, or vice versa.

Key concept

The time delays, and the strengths and directions of the reflected signals, are used to obtain information about the sizes, positions and natures of the surfaces and objects from which the waves were reflected.

Key concepts

Ultrasound is used for imaging inside the body because the waves can penetrate the human body, but some of the waves are reflected every time they meet an interface (boundary) between different media.

Ultrasound imaging provides a quick, safe, economical and mobile way of examining inside the body, especially when soft tissues are involved, but resolution is poor.

Generation and detection of ultrasound in medical contexts

Revised

Key concept

An alternating voltage across a piezoelectric crystal transducer makes it vibrate at the same frequency. This sends mechanical waves (ultrasound) into the surrounding materials.

An alternating voltage is induced across the same transducer when it receives back reflected ultrasound waves.

- Ultrasound waves are produced using the **piezoelectric effect** (see Figure 15.56).
- Vibrations are transferred from the transducer to the patient with a **gel** between them which eliminates air (more details below).
- When reflected waves are received back at the probe, oscillating voltages are produced and detected.
- The ultrasound waves are transmitted in pulses, with sufficient time between the pulses for the reflected waves to be clearly detected, as shown in Figure 15.57. Resolution is improved by having several complete ultrasound waves in each pulse.

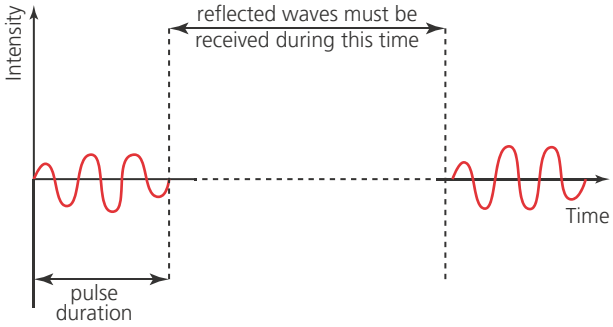


Figure 15.57

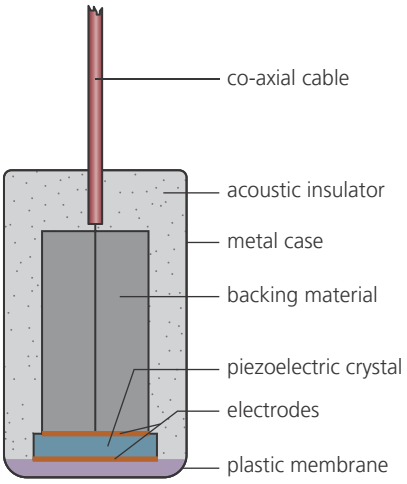


Figure 15.56

Acoustic impedance

Revised

- The term **acoustic** describes anything related to sound.
- As ultrasound waves travel through any medium, they become attenuated: their intensity is reduced because of scattering and absorption. The amount of attenuation varies with the medium and the frequency of the waves. The intensity decrease of a parallel beam can be described by the equation $I = I_o e^{-\mu x}$. (The same equation as used for the attenuation of X-rays.)
- The speed of an ultrasound wave through a medium, c, depends on its acoustic impedance and its density, ρ.
- The ratio of reflected intensity, I_r, to incident energy, I_0, at a boundary between two media of acoustic impedances Z_1 and Z_2 is given by the following equation (knowledge of which is not required in the IB Physics course):

$$\frac{I_r}{I_0} = \frac{(Z_2 - Z_1)^2}{(Z_2 + Z_1)^2}.$$

- The acoustic impedance of a particular medium increases with the frequency of the wave.

QUESTIONS TO CHECK UNDERSTANDING

60 Explain why ultrasound scans are not usually made of the brain or the lungs.

61 Under what circumstances would you expect ultrasound waves to change direction as they pass from one medium into another?

62 The range of frequencies used in ultrasound imaging is approximately 1–20 MHz. Estimate the longest wavelength used.

Key concepts

Acoustic impedance, Z, is a measure of the opposition of a medium to the flow of sound through it at a particular frequency. It depends on the speed of the wave and the density of the medium:

$Z = \rho c$

The units of acoustic impedance are $\text{kg m}^{-2}\text{ s}^{-1}$.

When an ultrasound wave meets a boundary between different media, the percentage of incident waves that are reflected depends on the difference between their acoustic impedances.

Solving problems involving ultrasound acoustic impedance, speed of ultrasound through tissue and air and relative intensity levels

QUESTIONS TO CHECK UNDERSTANDING

63 Consider Figure 15.57.

- **a** If the frequency being used was 5.0 MHz, what was the time duration of each pulse?
- **b** If the waves were being used to examine an organ which was 8.0 cm below the skin, what is the maximum pulse repetition frequency? (Assume speed of waves = 1550 m s^{-1}.)

64 a What is the acoustic impedance of muscle if its density is 1.08 g cm^{-3} and the speed of sound waves in the tissue is 1600 m s^{-1}?

- **b** The acoustic impedance of bone is about five times greater than muscle, and it has approximately twice the density. Estimate the speed of ultrasound in bone.

65 a Ultrasound is travelling from tissue of impedance 1.42 × 10^6 kg m^{-2} s^{-1} into liver of impedance 1.66 × 10^6 kg m^{-2} s^{-1}. Use the equation quoted above to determine the percentage of the wave intensity which is reflected.

- **b** Repeat the calculation for waves travelling in the opposite direction.
- **c** If at a higher frequency both impedances were increased by 20%, how would your answer to (a) change?

66 The attenuation of ultrasound in soft tissue (at 1 MHz) averages at 0.54 dB cm^{-1}.

- **a** What percentage of 1 MHz ultrasound waves are scattered and absorbed by travelling through 1 cm of soft tissue?
- **b** If the percentage scattered and absorbed is proportional to frequency, what is the attenuation per cm (dB) at a frequency of 2 MHz?

Expert tip

The terms *resistance* and *impedance* are also widely used in the topic of electricity. The opposition to the flow of a *direct* current through a conductor is called its resistance, but there are other factors opposing the oscillations of *alternating* currents (capacitance and inductance). The total opposition to the flow of ac is called impedance. The impedance of a conductor is not constant; its value varies with the frequency involved, as does acoustic impedance.

Explaining the choice of frequency and use of gel

- **Choice of frequency.** An ultrasound frequency of, for example 1.5 MHz, has a wavelength, $\lambda = 1.0$ mm in soft tissue. We know from Chapter 9 that diffraction effects and resolution depend on the ratio $\frac{\lambda}{b}$ (the smaller the better), where b is the size of the diffracting object. Although a wavelength of about 1 mm is smaller than the aperture of the transducer and the size of most of the objects that the waves will be used to examine, diffraction effects will be great enough to reduce the quality of the images. Resolution can be improved by *increasing* the frequency.
- However, attenuation of ultrasound will be reduced by *decreasing* the frequency.
- These two factors must be balanced against each other and the best frequency will depend on the type of scan being performed. For example, higher frequencies may be the better choice for locations which are just beneath the skin (when absorption is less important).
- **Use of gel**. The acoustic impedance of skin is approximately 5000 times greater than air. This means that nearly all the ultrasound waves incident on skin from air will be reflected. The use of a suitable gel between the transducer and the skin removes the air and overcomes this problem. The gel should have an acoustic impedance which is similar to that of skin and the surface of the transducer.

Key concepts

Resolution can be improved by *increasing* the ultrasound frequency. However, attenuation of ultrasound will be improved by *decreasing* the frequency. The best frequency to use depends on the type of scan.

A *gel* is used between the transducer and the skin in order to maximize the energy transferred.

Explaining the difference between A and B scans

Revised

- A medical **scan** is the common name given to the process of obtaining an image from inside the human body, for example, an **ultrasound scan** (or X-ray scan).

- 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. Figure 15.58 shows an example, but note that the angles have been distorted in order to show the various reflections clearly and the time scale is not regular.

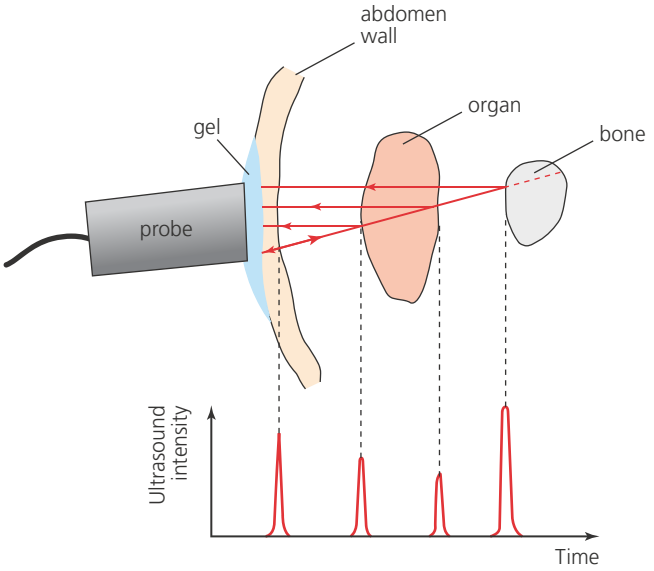


Figure 15.58

- **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 slightly different directions, often while the probe is moved to different positions (Figure 15.59).

Key concepts

The results of an *A-scan* are displayed as a graph of the amplitude of the reflected waves against time.

B-scans produce a real-time two-dimensional image, with brightness representing the intensity of the reflected waves.

Figure 15.59

QUESTIONS TO CHECK UNDERSTANDING

67 a Explain why the resolution seen on images obtained using ultrasound is much worse than the resolution achieved by the human eye.

b Why is resolution of ultrasound images improved by using higher frequencies?

68 Why are lower ultrasound frequencies used for obtaining images from deep inside a human body?

69 a Use the equation $\frac{I_r}{I_0} = \frac{(Z_2 - Z_1)^2}{(Z_2 + Z_1)^2}$ to explain why a gel is needed between the transducer and the skin when making an ultrasound investigation.
($Z_{air} = 4.1 \times 10^{-4}\,kg\,m^{-2}\,s^{-1}$, $Z_{skin} = 2.0\,kg\,m^{-2}\,s^{-1}$)

b Suggest a suitable value for the acoustic impedance of the gel.

70 Consider Figure 15.58.

a Explain why the amplitudes of the first three peaks are getting smaller and smaller.

b Explain why the last peak is larger than the others.

c If the organ shown in the figure is a kidney which has a density of $1050\,kg\,m^{-3}$ and an acoustic impedance of $1.62 \times 10^6\,kg\,m^{-2}\,s^{-1}$:

i What is the speed of ultrasound waves in the kidney?

ii What is the 'size' of the kidney if the delay between reflections received from its surfaces is 0.28 ms?

Medical imaging techniques (magnetic resonance imaging) involving nuclear magnetic resonance

Revised

- **Nuclear magnetic resonance** (NMR) is a process used to investigate internal structures. The object under investigation is placed in a strong uniform magnetic field and exposed to electromagnetic radiation of suitable frequency (usually radio frequencies, RF). The way in which the material responds to the radiation depends on the behaviour of the nuclei within it.
- **Magnetic resonance imaging** (MRI) is the use of NMR for medical diagnoses.

MRI scans

Revised

- In magnetic resonance imaging, radio waves (generated in external **RF coils**) are used to excite the protons of hydrogen atoms within the patient's body placed in a strong magnetic field. The pattern of the waves subsequently re-emitted by the protons is characteristic of the properties of the tissues involved. That is, analysis of the waves received back at the coils leads to knowledge of the locations and nature of the tissues from which they were emitted.
- Hydrogen atoms are found within molecules throughout the human body.
- Figure 15.60 shows a typical arrangement for an MRI scan. The patient must remain still during the process.

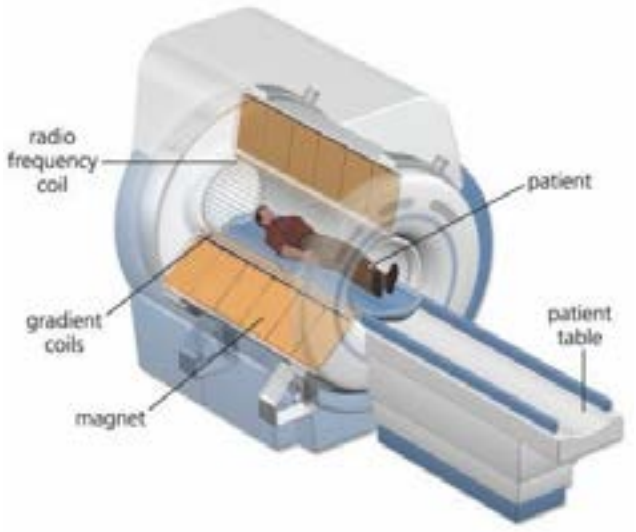


Figure 15.60

- All protons spin and this movement of charge creates a tiny magnetic field around each proton. However, under normal circumstances, all these spins are orientated randomly, so that there is no overall magnetic field that can be detected.
- However, if the material containing the protons is within a strong uniform external magnetic field, the protons will align with the field and rotate (**precess**) around it. Precession may be understood by considering a child's spinning top, as shown in Figure 15.61.

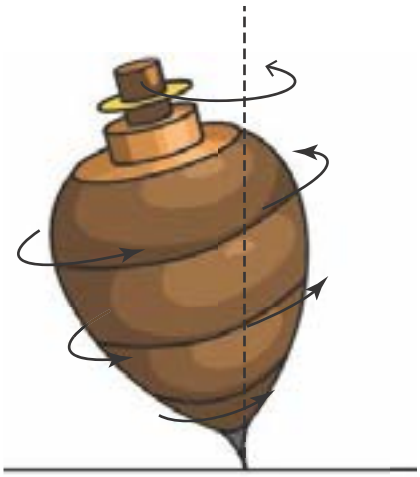


Figure 15.61

- The rate of proton precession is proportional to the strength of the magnetic field and is called the **Larmor frequency**. For example, in a field of strength 1 T, the Larmor frequency is 4.3×10^7 Hz. This frequency is in the radio wave (RF) section of the electromagnetic spectrum. Stronger magnetic fields and higher Larmor frequencies generally improve image quality.

Explaining the origin of the relaxation of proton spin and consequent emission of signal in NMR

- Energy can be transferred to the spinning protons by using an oscillating electromagnetic (RF) field of the same (Larmor) frequency. This is supplied from coils around the patient.
- The protons will be made to precess in phase by the RF field and some will have enough energy to 'flip' from being parallel to the field to the higher energy state of being anti-parallel to the magnetic field.
- **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.
- When the RF signal stops, the protons return to their equilibrium state within the magnetic field at a rate which depends on the type of medium in which they are located. This process is known as **relaxation** of proton spin.
- The radiation emitted by the relaxing protons can be detected by the same RF coils around the patient.

Explaining the use of gradient fields in NMR

- As stated above, protons are made to precess at a known rate by the presence of a very strong, uniform (primary) magnetic field. However, by having controllable additional (secondary) magnetic fields, each varying in a regular and known way in three direction (x, y and z), different parts of the patient can be made to resonate at different frequencies, thus allowing reconstruction of the three-dimensional distribution of protons. These are known as **gradient fields**.

Discussing the advantages and disadvantages of ultrasound and NMR scanning methods, including a simple assessment of risk in these medical procedures

- Both ultrasound and NMR are useful for obtaining images of soft tissue within the human body. They are also both considered to be risk-free procedures under most circumstances (unlike the small risks associated with the ionizing radiation used in X-rays and CT scans).

QUESTIONS TO CHECK UNDERSTANDING

71 Why is the use of MRI considered to be safer than the use of X-rays for obtaining images from within the body?

72 What is the Larmor frequency for protons in a magnetic field of strength 2.5 T?

73 Explain the origin of the RF waves received by the coils (which are used to construct the image).

74 What is meant by the term *gradient field* in an MRI scan?

75 Explain why MRI is considered to be an example of resonance.

76 Explain why MRI can be used to obtain images of the brain, but ultrasound cannot.

Expert tip

The very strong magnetic fields needed for MRI scans are not considered to be a risk to health (for most patients). Very large currents are needed in the coils which produce such strong fields and usually this is achieved by cooling the circuits to very low temperatures so that their electrical resistance is reduced to very low values (superconducting).

Key concept

When the protons of hydrogen atoms are placed in a strong uniform magnetic field they align with the field and *precess* at the *Larmor frequency*. This frequency depends on the strength of the magnetic field.

Key concepts

Energy can be transferred to the protons using an RF field oscillating at the Larmor frequency. This is a form of *resonance*.

After the RF field is removed, the protons return to their original state (*relax*) and emit RF radiation at a rate which depends on the material in which they are located.

Key concept

The use of *gradient fields* allows protons in different parts of the patient's body to resonate at different frequencies.

- *NMR* usually produces images with better resolution, sharpness and contrast. Much greater data is obtained and different planes and orientations can be selected for inspection after the process. Images can be obtained within and behind bone (for example, in the brain).
- The advantages of using *ultrasound* tend to be more practical: the equipment is much easier to use and less expensive. Images can be seen in real time and patients do not have to remain still for a long time in difficult surroundings.

NATURE OF SCIENCE

Risk analysis

All medical procedures involve a certain amount of risk to the patient. Passing various radiations through the body is an obvious example. It is a crucial part of the job of medical staff to be aware of the extent of the risks when making judgements or offering advice about possible courses of treatment. All patients and their circumstances are different, so there can be no certainty of outcome. Decisions need to be made on the balance of probabilities as determined by current scientific knowledge.

Option D 16 Astrophysics

16.1 Stellar quantities

Revised

Esssential 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.

Objects in the universe

Revised

- Most of this chapter is about the nature of stars, but many other types of objects can be identified in the universe.

Identifying objects in the universe: stars

- Large nebulae are the principal location for the formation ('birth') of stars.
- Every star has a 'birth', a 'lifetime' and a 'death' (explained in detail later). For most of its lifetime, a star emits radiation that has been transferred from nuclear fusion of the nuclei of gas atoms at its core.
- The closest star to Earth is, of course, the **Sun**.
- Many stars, called *binary stars*, exist in pairs.
- All stars are moving but, because they are at such enormous distances away, there is no motion obvious to observers on Earth, even over hundreds of years. (Although tiny shifts in position can be detected for some stars which are relatively close to Earth. This is called *stellar parallax* and it is discussed later.) Because of this lack of apparent relative motion, **star maps** can be drawn showing the relative positions of the stars. Figure 16.1 shows a star map for the Southern hemisphere night skies.
- As the Earth rotates, the view of the stars above our heads gradually changes during the night.
- Because the Earth is moving around the Sun, the star pattern that we see at midnight (for example) is slightly shifted every night (see Figure 16.2).
- The intensity of visible light in the day-time prevents the observation of the stars, although they can still be detected by radio telescopes.

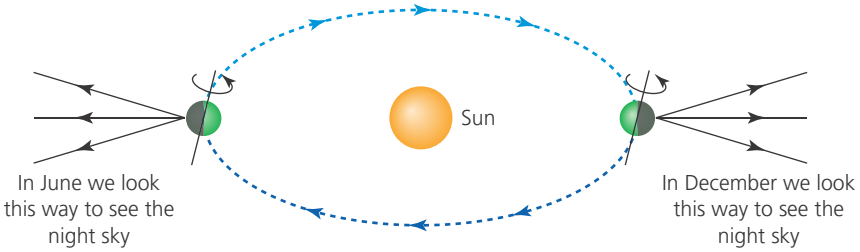


Figure 16.2

- Constellations are named after a well-known group of visible stars within that region (e.g. Orion). By naming the *constellation* in which a star is located, we know where to look for it. As seen from Earth, the stars of a constellation have relatively small *angular* separations, but it is possible, because of the three-dimensional nature of space, that a star could be nearer to a star in another constellation than to other stars in the same constellation.
- *Stellar clusters* move as a group of stars, rather than independently, because they are close enough for gravitational forces to bind them together. (Stellar means related to stars.) **Globular clusters** contain a very large number of

Key concepts

Nebulae are enormous diffuse 'clouds' of **interstellar matter**, which is mainly gases (mostly hydrogen and helium) and some dust containing traces of other elements.

A **star** consists of a very large mass of hot **plasma** (highly ionized gas) that has been pulled by gravity into a spherical shape.

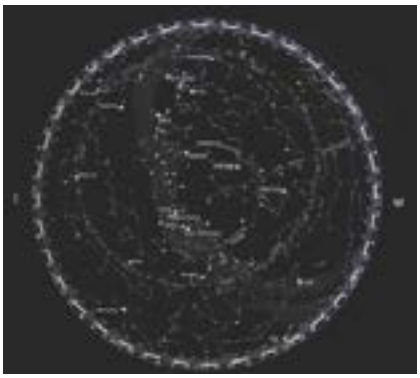


Figure 16.1

Expert tip

Two-dimensional star maps have obvious limitations in trying to represent three-dimensional space. An imaginary sphere around the Earth on which the apparent positions of the stars is marked is called a *celestial sphere*. (*Celestial* describes anything to do with space.)

Key concepts

Different regions on a star map (or celestial sphere) are known as **constellations**.

A **stellar cluster** is a collection of stars which were formed in the same nebula and move as a group within their galaxy.

stars in which gravitational forces, acting over a long period of time, have produced a combined spherical shape (globular). **Open clusters** have fewer stars and are younger. The cluster has a much less well-defined overall shape.

Identifying objects in the universe: galaxies

- Galaxies typically have many more stars than are found in stellar clusters, and the stars within them have more varied origins. Each galaxy rotates about its centre of mass.
- Some of the spots of light seen with a simple telescope or binoculars are distant galaxies, rather than individual stars. Galaxies are usually classified by their shapes: principally, elliptical, spiral and irregular.
- All the stars that we can see from Earth are situated within our own galaxy, which is called the **Milky Way**.
- Galaxies are usually grouped into clusters. Each cluster is approximately spherical in shape and may contain tens, hundreds or thousands of galaxies. (The Milky Way is in a cluster of about 50 galaxies called the Local Group.)
- Superclusters of galaxies are among the largest known structures in the universe (see Figure 16.3, in which each dot is a galaxy).

Key concepts

A **galaxy** is a group of billions of stars and other interstellar matter (including *dark matter*) bound together by gravity.

Gravitational forces result in galaxies coming together in groups called **clusters of galaxies**.

Clusters of galaxies are themselves usually located in even bigger groups, called **superclusters**.

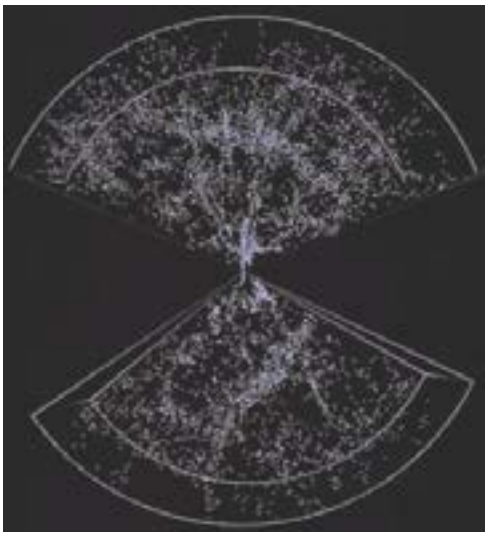


Figure 16.3

Identifying objects in the universe: planetary systems

- **Planets** move in **elliptical** paths with periods which depend on the mass of the star and the distance from the star.
- **Comets** are much smaller than planets with typically much longer periods and more elliptical paths. They are composed of dust and ice, and when they are close to the Sun (and therefore the Earth), they may become visible to us, and have a 'tail' of particles created by increased solar radiation (see Figure 16.4).

Key concept

Our **solar system** is an example of a **planetary system**: planets, comets and other objects orbiting a star (the *Sun*).

Figure 16.4

Expert tip

Some planets have one or more massive rocky objects orbiting around them. They can be described as natural satellites and they are called **moons**. Apart from the planets and comets, there are other smaller rocky objects orbiting the Sun. Most of these are located between the orbits of Mars and Jupiter, and they are called *asteroids*.

Common mistake

The word *cluster* usually means a group of similar things relatively close together, but the widespread use of this term can cause confusion because it is applied to groups of stars within a galaxy, as well as groups of galaxies. The term *galactic cluster* may be best avoided because it usually refers to stellar clusters within a single galaxy.

The nature of stars

Revised

- The dominant nuclear fusion in stars is the fusion of hydrogen into helium, which can be simplified as:

$$4\,{}^{1}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + 2\,{}^{0}_{1}\mathrm{e} + \text{neutrinos and photons.}$$

- 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. The fusion of heavier elements generally occurs later in the lifetime of stars.

Qualitatively describing the equilibrium between pressure and gravitation in stars

- **Stellar (hydrostatic) equilibrium** is represented in Figure 16.5. A star can remain in equilibrium for a very long time. During this time, it is said to be on the **main sequence** (this important term is explained later). The rate of energy transfer from nuclear fusions equals the rate at which energy is radiated away from the star.
- Eventually, the supply of hydrogen will be reduced enough that 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 (more details later).

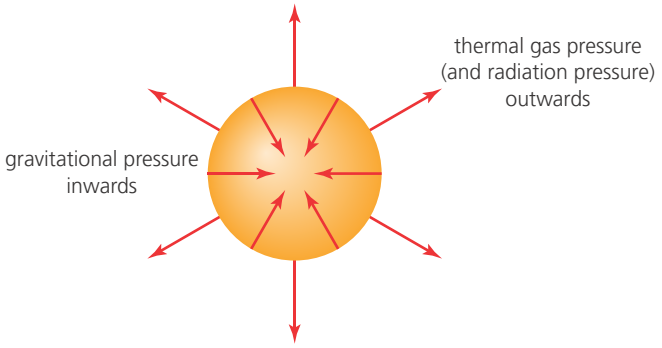


Figure 16.5

Binary stars

- Figure 16.6 provides an artistic impression of a pair of binary stars.
- The period of their orbital motion is very useful information because it enables astronomers to calculate their mass. (Larger masses orbit more slowly.)
- Most binary stars are relatively close together and too far away to be directly detected as two separate stars, but detailed observation of the radiation received from them can provide evidence for their binary nature: (1) the intensity will change periodically if one star passes 'in front of' (eclipses) the other; (2) The radiation will have opposite *Doppler shifts* if they are moving rapidly in opposite directions (explained later).

Key concept

Over a very long period of time in a nebula, gravity pulls gas atoms closer together and eventually they can gain very high kinetic energies if the overall mass is large (so that the temperature becomes extremely high – millions of kelvin). At the centre, 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 begins to happen on a large scale, it is called the *birth* of a star.

Key concept

A star can remain in *equilibrium* for a long time because the **gravitational pressure** inwards is balanced by **thermal gas pressure** and **radiation pressure** outwards.

Key concept

In a **binary star system**, two stars orbit their common centre of mass.

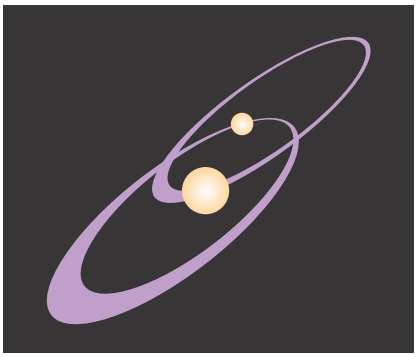


Figure 16.6

QUESTIONS TO CHECK UNDERSTANDING

1 Why can the Moon and the planets (and the Sun) not be located on star maps?

2 Discuss why some teachers may like to use an inflated balloon as an analogy for a star.

3 The power emitted by the Sun is 3.85×10^{26} W. Estimate how many helium atoms are fused every second in the Sun.

4 Distinguish between a cluster of stars and a galaxy.

5 a Give two reasons why comets are only seen rarely from Earth.
b Make a sketch of the Earth's orbit around the Sun.
c Add to your sketch a possible path for a comet.
d What causes the visible 'tail' seen on some comets?

6 Why does the search for life in other parts of the universe concentrate on planets?

7 Kepler's third law (which need not be remembered) states that the average radius of a planet's orbit, r, and its period, T, are related by $\frac{r^3}{T^2} = \frac{GM}{4\pi^2}$, where M is the mass of the Sun. What would be the period of a planet which orbited the Sun at a distance 50 × greater than the Earth's orbit?

8 a Explain with the help of a diagram, why the intensity of the radiation received from binary stars can vary periodically.
b Why is this variation only observed from some, but not all, binary systems?

Astronomical distances

Revised

- The distance to stars is obviously vital knowledge in astronomy. For stars that are relatively close to the Earth, their distance away can be estimated by using trigonometry in the *stellar parallax* method (explained below). However, for most stars we need to use other methods. These usually involve identifying a star of a kind which has a known power and relating that power to the intensity of radiation received on Earth. Alternatively, measurements of *red-shift* can be used to determine the distance to distant galaxies. These methods are explained later.
- The enormous magnitudes of distances in astronomy (and their differences) is very difficult for us to comprehend. Table 16.1 is meant to represent that range, but there is no need to remember any specific details.

Table 16.1

Height of the lowest satellite orbit around the Earth	≈ 200 km
Distance to the Moon	≈ 384 000 km
Distance to the Sun	≈ 150 000 000 km
Distance to the 'edge' of the solar system	≈ 6 000 000 000 km
Distance to the nearest star (other than the Sun)	≈ 9 500 000 000 000 km
Distance to the edge of the Milky Way	≈ 300 000 000 000 000 000 km
Distance to the nearest galaxy (Andromeda)	≈ 24 000 000 000 000 000 000 km
Diameter of a galactic supercluster	≈ 5 000 000 000 000 000 000 000 km
Distance to the edge of the observable universe	≈ 900 000 000 000 000 000 000 000 km

Using the astronomical unit, light year and parsec

- The metre and the kilometre are obviously inconveniently small units for measuring astronomical distances. So astronomers have developed the use of several other (non SI) units for measuring distance. These units are shown on the right and compared in Table 16.2.

Key concepts

The **light year** (ly) is equal to the agreed distance travelled by light in vacuum in one year. **1 light year (ly) = 9.46×10^{15} m**

The **astronomical unit** (AU) is equal to the agreed mean distance between the Earth and the Sun. **1 astronomical unit (AU) = 1.50×10^{11} m**

The **parsec** (pc) is equal to the distance to a star which has a parallax angle of one arc-second. **1 pc = 3.26 ly**

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

Approximate distances are shown in Figure 16.7.

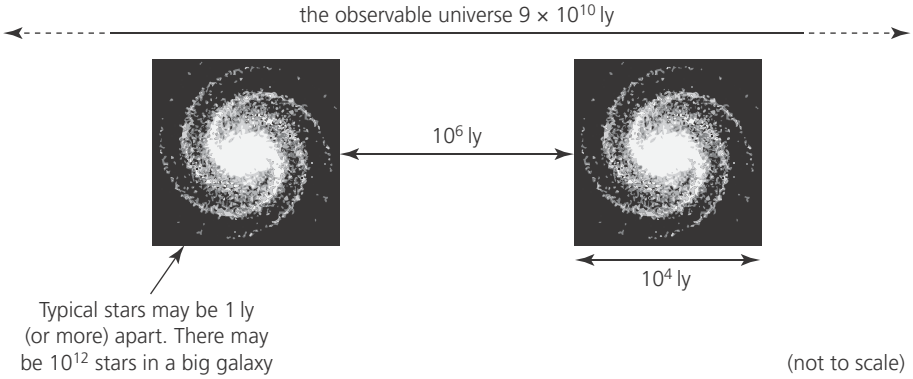


Figure 16.7

Stellar parallax and its limitations

- **Parallax** is the displacement in the apparent position of an object (compared to its background) when it is viewed from different positions.
- When we describe the position of a star in the sky, we do so by comparing it to the positions of other stars apparently near to it in the same constellation on a star map. For most stars this position does not change, but for stars which are relatively close to Earth, this apparent position changes *very slightly* during the year. This is called **stellar parallax** and the effect can be used to determine the distance to the star. The greatest difference occurs between the apparent positions of the star when viewed at times which are 6 months apart, as shown in Figure 16.8.
- For the sake of clarity, the size of the parallax angle has been *greatly* exaggerated in Figures 16.8 and 16.9. In practice, parallax angles are fractions of **arcseconds**. (There are 3600 arcseconds in an angle of one degree.)

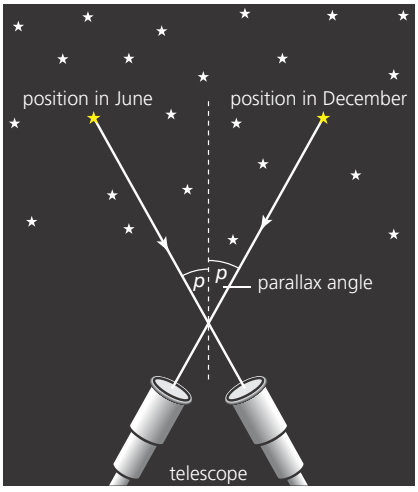


Figure 16.8

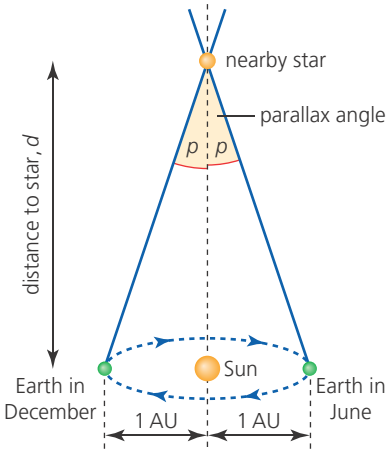


Figure 16.9

Describing the method to determine distance to stars through stellar parallax

- The measurements needed to determine the distance, d, to a star using stellar parallax are shown in Figure 16.9.
- Parallax angle, $p = \frac{1}{d\ (AU)}$ (p is in radians)
- For example, the parallax angle for the star Proxima Centauri (the nearest star to Earth after the Sun) is 0.772 arcseconds. This is about 2.1×10^{-4}° or 3.7×10^{-6} rad. Using the previous equation, $d = (2.7 \times 10^5\ AU) = 4.0 \times 10^{16}$ m (which is 4.2 ly).
- However, rather than using this equation to calculate the distance in metres, kilometres or light years, stellar distances can be directly quoted in the unit of parsecs (pc):

$$d\ (parsec) = \frac{1}{p\ (arcsecond)}$$

- This is an inverse relationship: larger parallax angles mean smaller distances. A star which has a parallax angle of 0.50 arcseconds is 2 pc away, while a star with a parallax angle of 0.25 arcseconds is 4 pc away, etc.
- Returning to the previous example, the distance to Proxima Centauri $= \frac{1}{0.772} = 1.30$ pc.
- The unit parsec is used generally in astronomy, not just for examples of stellar parallax.
- For stars further away than about 100 parsecs, the stellar parallax method cannot be applied because the parallax angle is too small (less than 0.01 arcseconds) to measure accurately.

Key concepts

The distance to a 'nearby' star (within 100 pc) can be determined by using geometry and its stellar parallax angle.

The distance to a star which has a stellar parallax angle of 1 arcsecond is called one parsec (pc).

QUESTIONS TO CHECK UNDERSTANDING

9 **a** What is the distance to the edge of the visible universe in light years?
b What is the approximate diameter of the solar system in AU?
c How far away is the galaxy Andromeda in pc?

10 The parallax angle to a star is measured to be 0.22 arcseconds. Determine the distance to this star in
a parsec
b kilometres.

11 The star Procyon is 11.4 ly from Earth.
a How far is this in parsec?
b Determine the parallax angle to this star.
c Would it be possible to use the method of stellar parallax to determine its distance from Earth?

12 **a** What is the approximate diameter of the Milky Way in parsec?
b Explain why the distance to most stars in the Milky Way cannot be determined by stellar parallax.

Luminosity and apparent brightness

Revised

- If the size of a star and its surface temperature are known, its luminosity, L, (power) 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 the star behaves like a *perfect black body* and has an emissivity of 1. (Reminder: the surface area of a sphere $= 4\pi r^2$.)
- For example, the star Betelgeuse has a radius of 8.8×10^8 km and a surface temperature of 3590 K. The highlighted equation in the box on page 115 can be used to show that its luminosity is approximately 9×10^{31} W.

- In practice, this would be an unusual calculation to make because astronomers are much more likely to use a known luminosity of a star to determine its properties (than the other way around).
- The intensity of radiation received at the Earth from a star (before the effects of the Earth's atmosphere are taken into consideration) is called the *apparent brightness*, b, of the star, and it depends only on the luminosity of the star and its distance away. This assumes that no energy is absorbed in interstellar space (see comment below).
- Understanding the relationship between luminosity and apparent brightness is very important in the study of astronomy (see Figure 16.10).

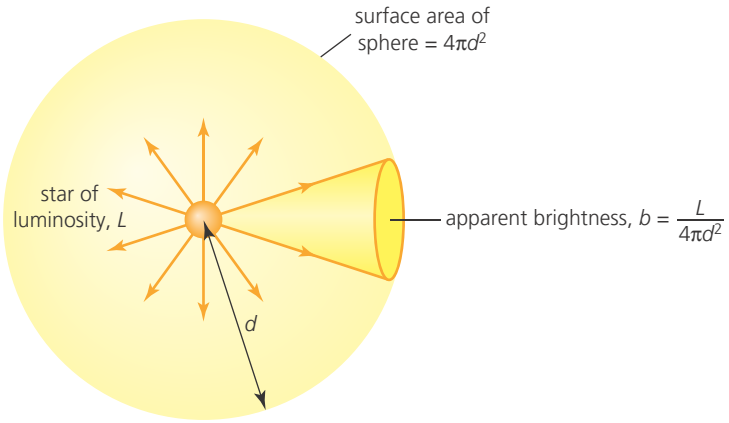


Figure 16.10

- Apparent brightness can be calculated from $b = \frac{L}{4\pi d^2}$, 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 may lead to inaccuracies if this equation is used to determine distances.
- For example, Betelgeuse is a distance 6.1×10^{18} m from Earth. The preceding equation can be used to show that its apparent brightness on Earth is about 2×10^{-7} W m^{-2}.
- It should be clear that the apparent brightness of a star is the primary measurement made on it (together with its direction). With this information, the equation $b = \frac{L}{4\pi d^2}$ can then be used to determine the luminosity of a star if its distance from Earth is known.

Solving problems involving luminosity, apparent brightness and distance

QUESTIONS TO CHECK UNDERSTANDING

13 Use the data above to confirm that
- **a** the luminosity of the star Betelgeuse is approximately 9×10^{31} W, and
- **b** its apparent brightness is about 2×10^{-7} W m^{-2}.

14 The star Sirius has a surface temperature of 9940 K and a luminosity of 9.8×10^{27} W. Estimate its radius.

15 When using $b = \frac{L}{4\pi d^2}$ to calculate the distance to a star of known luminosity, explain why the result may be an overestimate.

16 The apparent brightness of our Sun is 1360 W m^{-2} (the 'solar constant', as used in Chapter 8), and its luminosity is 3.85×10^{26} W. How far away is the Sun?

17 The star Altair is 11 times more luminous than the Sun. It distance from Earth is 17 ly. Determine its apparent brightness.

Key concepts

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.

$L = \sigma A T^4$

The **apparent brightness**, b, of a star is the intensity received from it at the Earth. It is measured in W m^{-2}.

$b = \frac{L}{4\pi d^2}$

Expert tip

For historical reasons, astronomers usually refer to the *magnitudes* of stars (rather than their brightness). This is an (initially somewhat confusing) *visual* brightness scale which relies on comparing the brightness of a star to other stars. Stellar magnitudes are not needed in the IB Physics course.

NATURE OF SCIENCE

Reality

The principles of physics that have been developed on Earth over hundreds of years have been applied with enormous success in recent years to developing an ever-increasing knowledge of the universe as a whole. Astronomers now have an understanding that not many years ago many would have believed to be impossible. All of this has been led by enormous advances in observational technologies.

16.2 Stellar characteristics and stellar evolution

Revised

Essential ideas: A simple diagram that plots the luminosity versus the surface temperature of stars reveals unusually detailed patterns that help us understand the inner workings of stars. Stars follow well-defined patterns from the moment they are created out of collapsing interstellar gas, to their lives on the main sequence and to their eventual death.

Stellar spectra

Revised

- Figure 16.11 can be used to explain why stars have appearances which have slightly different colours.

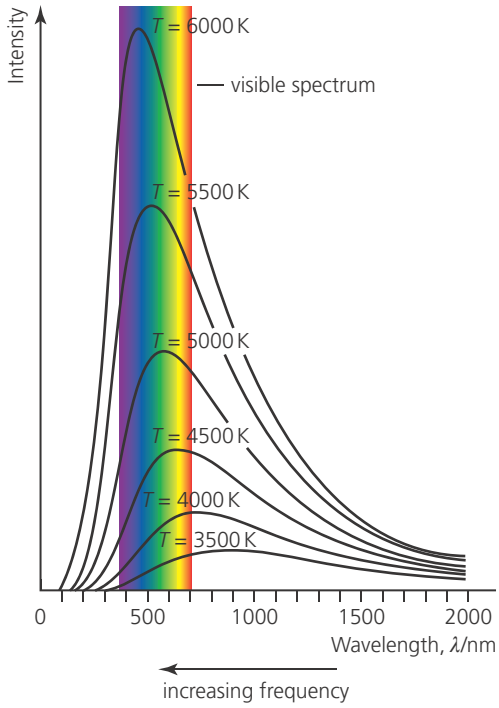


Figure 16.11 The black-body spectra emitted by stars with different surface temperatures.

Key concept

Stars can be considered to be black bodies and the continuous spectra emitted represented by intensity–wavelength graphs for different surface temperatures.

Explaining how surface temperature may be obtained from a star's spectrum

- From Figure 16.11, we can see that the wavelength at which the maximum intensity is radiated away from a star's surface depends on its temperature. This is represented by *Wien's displacement law* (Chapter 8).

Key concept

Wien's displacement law: $\lambda_{max} T = 2.9 \times 10^{-3}\,\text{mK}$ can be used to calculate the surface temperature, T, of a star if the wavelength at which the maximum intensity is received can be measured.

- For example, from Figure 16.11, we see that if the spectrum from a star has its peak intensity at about 580 nm, its surface temperature is approximately 5000 K, which can be confirmed by using Wien's displacement law. Such a star is emitting more radiation at longer visible wavelengths (than shorter wavelengths) and its overall appearance will be white-yellow in colour.

Explaining how the chemical composition of a star may be obtained from the star's spectrum

- Radiation emitted from the very hot core of a star has to pass through the cooler outer layers and some wavelengths of the continuous spectrum are absorbed and then re-emitted in random directions by the elements present.
- The absorption spectra received can be displayed by passing the light from the star through a diffraction grating or prism, as described and explained in Chapter 7, Section 7.1.
- As an example, Figure 16.12 shows the absorption spectrum of the Sun from which elements like helium and hydrogen (and others) can be identified.
- Figure 16.13 indicates how the continuous intensity–wavelength graph of radiation from a star with a surface temperature of about 5000 K (similar to the Sun and shown in Figure 16.11) is changed after passing through its outer layers.

Key concept

The elements present in the cooler outer layers of a star can be identified from the *absorption spectrum* received from the star.

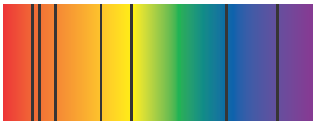


Figure 16.12

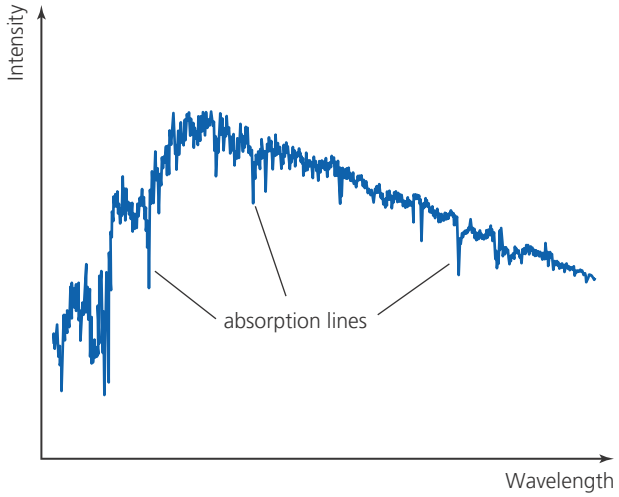


Figure 16.13

QUESTIONS TO CHECK UNDERSTANDING

18 Use information from Figure 16.11 to describe the overall appearance of a star with a surface temperature of 3500 K.

19 a At what wavelength is the maximum radiation intensity emitted from a star with a surface temperature of 5500 K?

b What is the colour of this wavelength?

20 What is the approximate temperature of a surface which emits radiation which has its maximum intensity at a frequency of 3×10^{14} Hz?

21 Most stars have surface temperatures in the range 2500 K to 40 000 K. Compare the appearances of stars at these two extremes.

22 a Explain the difference between

i a continuous spectrum and a line spectrum,

ii an absorption spectrum and an emission spectrum.

b Outline the process by which the intensity of light of particular wavelengths emitted by a star (in the direction of the Earth) is reduced in the outer layers of the star.

Main sequence stars

Revised

- Stable stars which are transferring energy because of the nuclear fusion of hydrogen to helium in their cores are described as **main sequence stars**. Most of the stars in the universe are main sequence stars.

Effect of mass on main sequence stars

- The most significant basic difference between main sequence stars is simply their mass.
- The mass of a star also controls what happens to it after the supply of hydrogen is reduced and it leaves the main sequence. This is explained in greater detail later.

Key concepts

Stars which are formed from greater masses will have greater gravitational forces pulling them together. This will result in higher temperatures at their core and much greater rates of nuclear fusion.

Mass–luminosity relation for main sequence stars

- For main sequence stars, the relationship between mass and luminosity is shown in Figure 16.14 (note the logarithmic scales). It is approximated mathematically by the equation $L \propto M^{3.5}$, which may be expressed as $\frac{L}{M^{3.5}} = \text{constant}$. This equation shows that relatively small differences in mass can have considerable effects on the luminosity (and lifetime) of a main sequence star.
- For example, if star A has twice the mass of star B, the equation above shows that star A will be about 11 times more luminous. So, it is transferring energy from nuclear fusion at 11 times the rate, but only has twice the mass. Of course, star A will come to the end of its main sequence lifetime much quicker than the less massive star B. More details are included in Section 16.4 (HL).

Key concept

More massive main sequence stars will also have greater sizes, greater surface temperatures and greater luminosities, but they will have shorter lifetimes because their supply of hydrogen will be fused much more quickly.

$L \propto M^{3.5}$

Applying the mass–luminosity relation

QUESTIONS TO CHECK UNDERSTANDING

23 Explain why more massive main sequence stars are:
- **a** more luminous,
- **b** bluer in colour.

24 Star X has twice the luminosity of Star Y. They are both main sequence stars.
- **a** What is the ratio of their masses?
- **b** Which star has the shorter lifetime on the main sequence?

25 The star Ursae Majoris is about 1.48 times more luminous than the Sun.
- **a** Estimate the mass (kg) of this star (mass of Sun = 2.0×10^{30} kg).
- **b** What assumption did you make when answering (a)?

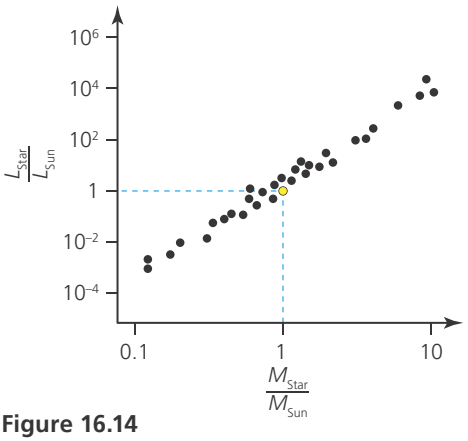


Figure 16.14

Hertzsprung–Russell diagram

Revised

- Luminosity and surface temperature are the two properties of any star which are easiest to determine. They are used on the axes of a diagram (HR), which is used to compare different stars and look for the existence of any possible patterns.
- The stars are *not* distributed randomly on a HR diagram (as some astronomers originally expected). Clear patterns and groups can be seen.

Key concept

The **Hertzsprung–Russell (HR) diagram** is a common way of locating different stars on the same chart (see Figure 16.15).

However, note that the temperature scale is reversed and both scales are logarithmic.

Sketching and interpreting HR diagrams

- The axes of the diagram are luminosity and temperature (reversed), but it is important to note the variations are so large that logarithmic scales are used.

- Figure 16.15 also attempts to represent the colours of the stars which, as has already been explained, is dependent on their surface temperature.
- We know that the luminosity of a star depends on its mass (and radius). The sizes of different stars can be compared on the HR diagram by using *lines of constant radius*.

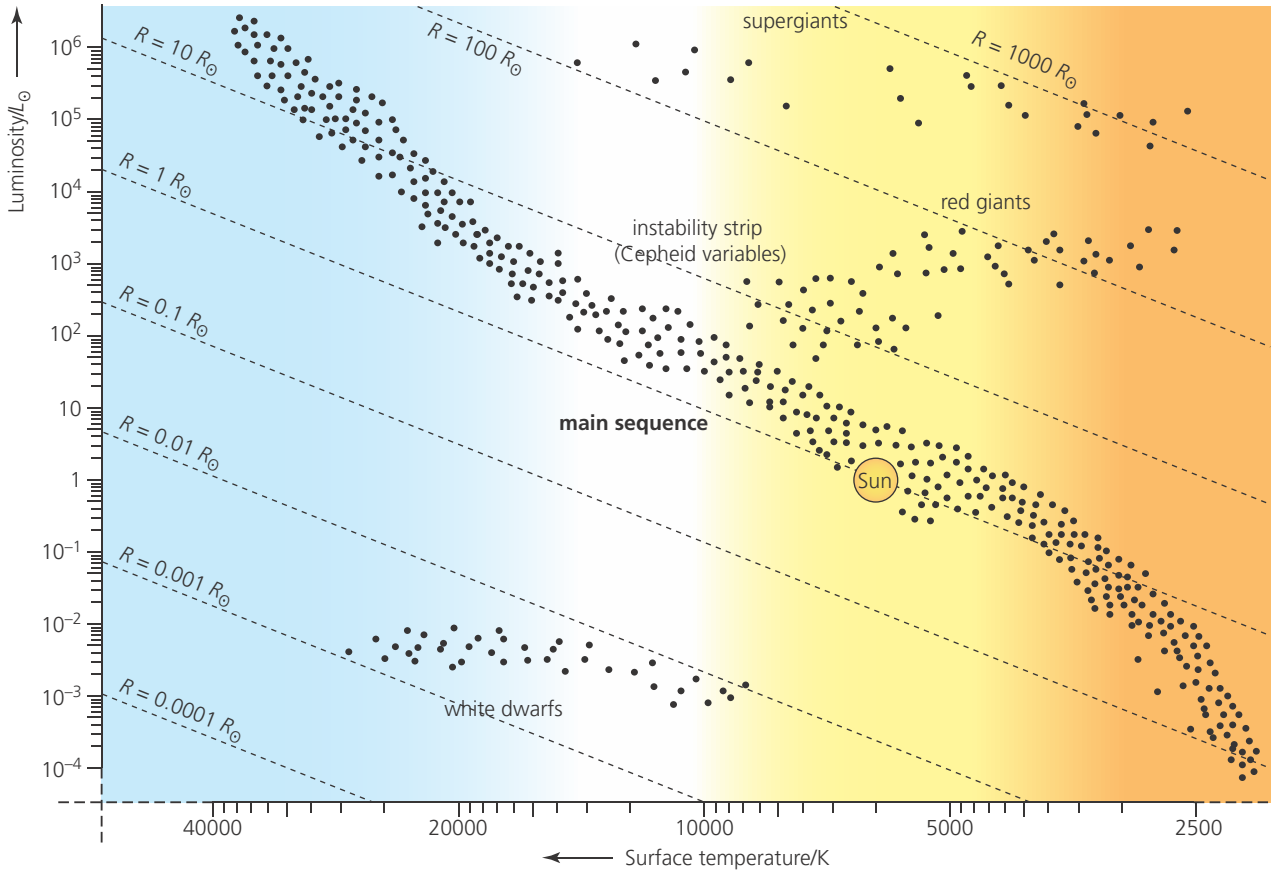


Figure 16.15

Identifying the main regions of the HR diagram and describing the main properties of stars in those regions

- The majority of stars (about 90%) are located somewhere in a group located around a diagonal line from bottom right to top left of the HR diagram. These are the *main sequence stars*. As already mentioned, moving from the bottom right to the top left, the main sequence stars increase in size, temperature and luminosity.
- Other types of non-main sequence stars, like *red giants*, *white dwarfs* and *supergiants* can be located in other parts of the HR diagram. Their names should indicate their positions on the HR diagram.
- After finishing their time on the main sequence, many stars have an interim period when they have various instabilities, often resulting in changes in their luminosity over (surprisingly) short time periods. Such stars are found in the **instability strip**. The stars in this region are *pulsating variable stars*, for example *Cepheid variables*.
- Further details about non-main sequence stars: red giants, white dwarfs, supergiants and Cepheid variables, are provided below.

Expert tip

The HR diagram is often drawn with *stellar class* (an indication of colour) on the horizontal axis, and *stellar magnitude* (a measure of luminosity) on the vertical axis. However, the IB Physics course does not need knowledge of these scales.

Key concept

The dominant feature of the HR diagram is the main sequence of stars running from bottom right to top left.

Other important regions of the diagram include, red giants and red supergiants, white dwarfs and the instability strip.

QUESTIONS TO CHECK UNDERSTANDING

26 Locate a star on the HR diagram which has a luminosity which is approximately 500 times less than the luminosity of the Sun, but has a hotter surface. Use the diagram to estimate the star's:

- **a** surface temperature,
- **b** colour,
- **c** radius.
- **d** What name do we give to this kind of star?

27 a Use the HR diagram to compare the luminosities of the brightest and dimmest main sequence stars.

- **b** Estimate the ratio of the masses of these stars.

28 There is a kind of star known as a *red dwarf*.

- **a** Where would these stars be located on the HR diagram?
- **b** Suggest why red dwarf stars have very long lifetimes.

Cepheid variables

Revised

- **Cepheid variables** are a particularly important kind of pulsating variable star. These stars are used to determine the distances from Earth to the galaxies in which they are located. As such, they are known as '**standard candles**'.
- Cepheid variables are so useful because the variations in their luminosity occur in typical times of only a few days (see Figure 16.16) and, most importantly, the periods can be directly related to the maximum (or average) luminosity of the star, as shown in Figure 16.17. Note the logarithmic scales on this **period–luminosity relationship** and that luminosity is given in multiples of the Sun's luminosity.

Key concept

Cepheid variables are pulsating stars. The period of their luminosity cycles can be used to determine their luminosity.

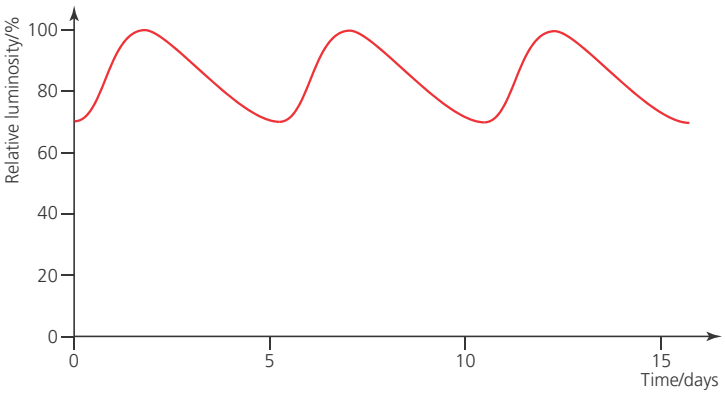


Figure 16.16

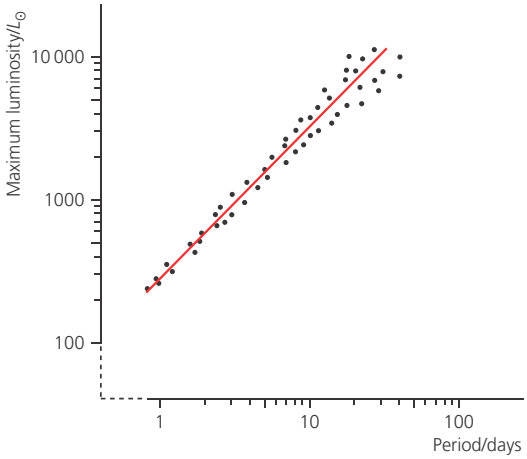


Figure 16.17

Determining distance using data on Cepheid variables

- As an example, the variation shown in Figure 16.16 has a period of approximately 5 days, and from Figure 16.17, we can see that this corresponds to a luminosity about 1500 times greater than the Sun, that is, $L \approx 6 \times 10^{29}$ W.
- 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.

Key concept

A graph of the *period–luminosity relationship* can be used to determine the luminosity of a Cepheid variable star which has the observed period. Then, if the apparent brightness of the star, b, has been measured, $b = \frac{L}{4\pi d^2}$ can be used to determine the distance, d, to the star and the galaxy in which it is situated.

Describing the reason for the variation of Cepheid variables

- The luminosity cycles shown in Figure 16.16 correspond to significant changes in the size of the Cepheid variable as its outer layers expand or contract.
- When the luminosity is at its lowest, the outer layers have an increased proportion of highly ionized helium. Under these conditions, the outer layers are less transparent and some photons are unable to escape. The photons are absorbed, energy is transferred and the star expands. The expansion causes the outer layers to cool, the proportion of ionized helium decreases, so that the outer layers become more transparent and the luminosity increases. Gravitational forces oppose the expansion, and the star oscillates under these competing effects.

QUESTIONS TO CHECK UNDERSTANDING

29 Explain the term 'standard candle'.

30 Describe the relationship between the maximum luminosities of Cepheid variables and the periods of their oscillations.

31 The luminosity of a certain Cepheid variable varies with a period of ten days.

a Use Figure 16.17 to estimate the maximum luminosity of the star in watts (luminosity of the Sun = 3.85×10^{26} W).

b Determine a value for the distance (in pc) to this star if its maximum observed brightness on Earth is 2.2×10^{-6} W.

32 Suggest why there are significant uncertainties in the measurement of distances using Cepheid variables.

Describing the evolution of stars off the main sequence

Revised

- Eventually the amount of hydrogen in the core of a main sequence star reduces to a level such that the amount of energy released is unable to resist *gravitational collapse*. This is the beginning of the end of the star's main sequence lifetime.
- Gravitational energy is transferred to particle kinetic energy and there is a relatively rapid rise in temperature, which is sufficient to start further fusion of hydrogen outside the core. The star is becoming a *red giant*.

Red giants

Key concepts

When the supply of hydrogen in the core of a main sequence star has been reduced to the point where equilibrium can no longer be sustained, the core begins to contract due to gravity, although there is still plenty of hydrogen outside the core. The contraction results in higher temperatures and more hydrogen can then fuse to form helium in a 'shell' around the core.

There is an enormous increase in the amount of energy released, and in the star's luminosity and size.

Despite the greater luminosity, the much greater surface area of the star results in a lower surface temperature, and its colour changes to become slightly red: the star becomes a red giant (or a red supergiant).

- The greater size and reduced *surface* temperatures of stars after they have finished their lifetime as main sequence stars (as described above) explains their name: **red giants** or **red supergiants**.
- For example, a red giant may be 100 times bigger and 2000 times more luminous than the main sequence star from which it evolved. The equation $P = e\sigma AT^4$ can be used to show that the surface temperature of the red giant will be about $\frac{2}{3}$ of the temperature of the main sequence star (e.g. about 3300 K instead of 5000 K).

- Depending on the mass (and therefore temperature) of the red giant star, nuclear fusion of heavier elements may be possible later. *Nucleosynthesis* is discussed further in Section 16.4 (HL).

Describing the role of mass in stellar evolution

- *Red giant stars* evolve from main sequence stars of mass less than approximately eight solar masses. More massive main sequence stars evolve into *red supergiants*. It is important to distinguish between these two because they evolve in different ways, as shown in Figure 16.18. The numbers in the figure represent the masses of the stars as multiples of the solar mass.

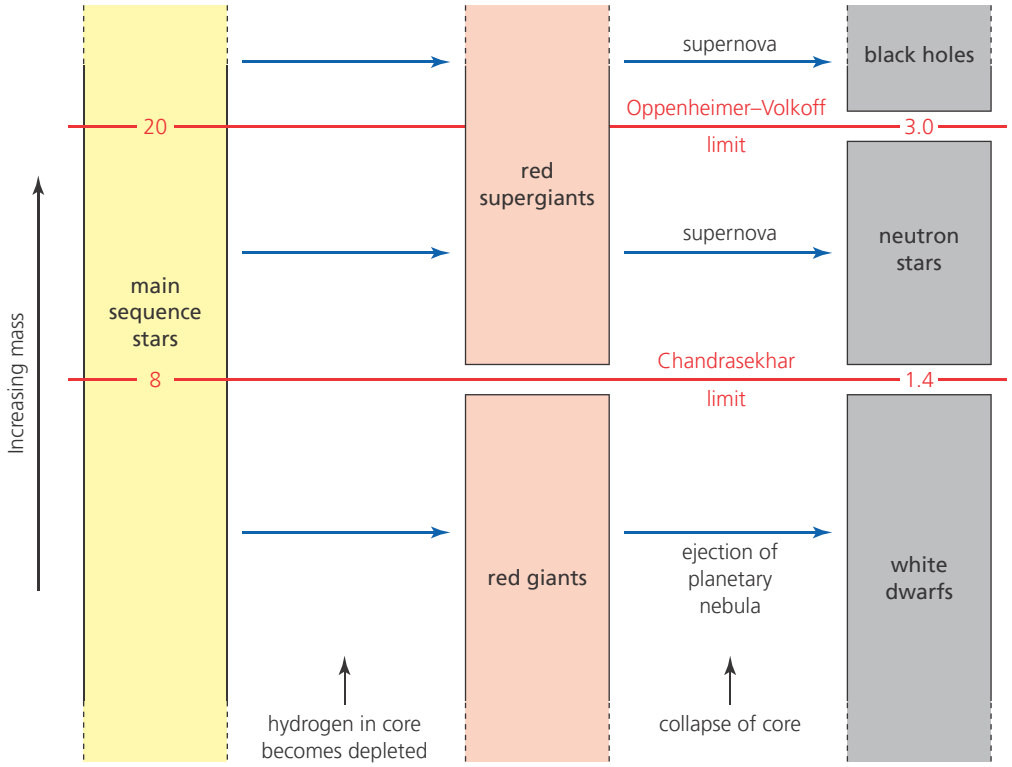


Figure 16.18

White dwarfs

- When nuclear fusion stops, the core of a red giant star collapses and the increase in temperature results in the ejection of its outer layers into the region around the star. This is known as a **planetary nebula**. (This misleading name is based on early descriptions, and it should be noted that *planetary nebulae* are nothing to do with planets or the vast nebulae in which stars are born.)
- The remaining core of the star becomes known as a **white dwarf** star.
- A process known as **electron degeneracy pressure** prevents the white dwarf collapsing further, so that this kind of star will remain stable for a long time. Electron degeneracy pressure requires a quantum physics explanation which is not required in the IB Physics course.
- Because of their small size white dwarf stars have low luminosities, but their surface temperatures are relatively high, so that they appear white (rather than the red/yellow colour of the red giant from which it was formed).
- The core of a star with a greater mass will be hotter and evolve in a different way, as described below.

Key concepts

When the nuclear fusion of hydrogen stops in the core of a red giant (which had an original mass *less* than 8 × solar mass), it collapses to become a *white dwarf star*. At the same time a *planetary nebula* is ejected.

Electron degeneracy pressure prevents the further collapse of the white dwarf.

Chandrasekhar limit

- The **Chandrasekhar limit** is the maximum mass of a white dwarf star (= 1.4 × solar mass).

Key concepts

It is the mass of a collapsed star which determines whether a white dwarf, a neutron star or a black hole is formed. These masses are specified in the *Chandrasekhar limit* and the *Oppenheimer–Volkoff limit*.

Neutron stars and black holes

- In the more massive red supergiants, electron degeneracy pressure is not great enough to prevent further collapse of the core and the resulting nuclear changes produce a massive explosion called a **supernova**. This results in either a *neutron star* or a *black hole*, depending on the mass of the red supergiant.
- A **neutron star** predominantly consists of tightly packed neutrons. It has a very small radius (typically 15 km) and extremely high density. Neutron stars are very hot.
- A neutron star can remain stable for a long time because of a process known as **neutron degeneracy pressure**. Neutron degeneracy pressure requires a quantum physics explanation that is not required in the IB Physics course.
- A **black hole** is a region of space where matter has become so compressed that the force of gravity is strong enough to prevent the emergence of electromagnetic radiation, including light. The presence of a black hole can be detected by its effect on other matter and radiation.

Key concepts

When the nuclear fusion of hydrogen stops in the core of a red supergiant (formed from a star which had an original mass *more* than 8 × solar mass), it collapses to form a supernova.

If the original mass of the star was between 8 × and 20 × solar mass, a *neutron star* is formed from the supernova. *Neutron degeneracy pressure* prevents the collapse of a neutron star.

If the original mass of the star was greater than 20 × solar mass, a *black hole* is formed from the supernova.

Oppenheimer–Volkoff limit

- The **Oppenheimer–Volkoff limit** is the maximum mass of a neutron star (≈ 3 × solar mass).
- If the core after a supernova has a mass of less than the Oppenheimer–Volkoff limit, it will contract to become a neutron star.
- If the mass is greater than the Oppenheimer–Volkoff limit, neutron degeneracy pressure is not great enough to resist the gravitational forces collapsing the core even more, and a black hole will be formed.

Stellar evolution on HR diagrams

Revised

- Figure 16.19 graphically represents the three different outcomes for main sequence stars which were described in the last section.

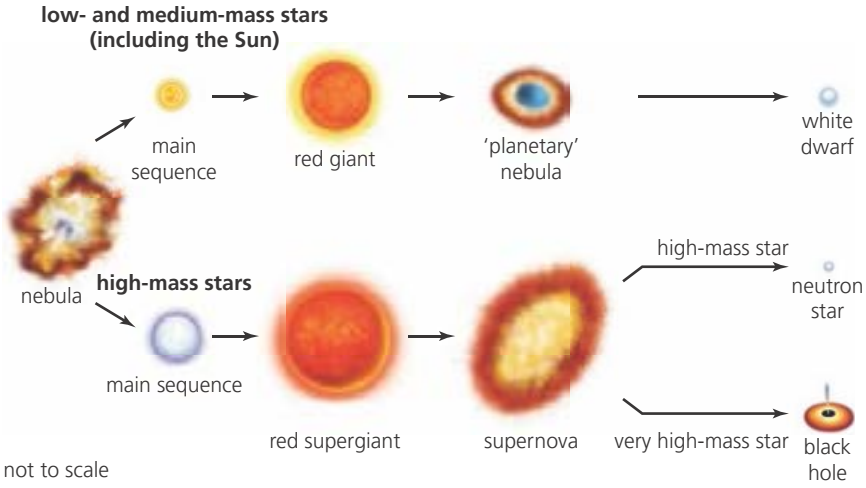


Figure 16.19

Sketching and interpreting evolutionary paths of stars on an HR diagram

The changes to stars after they leave the main sequence (like those shown in Figure 16.19) can also be traced on a HR diagram. Examples are shown in Figure 16.20.

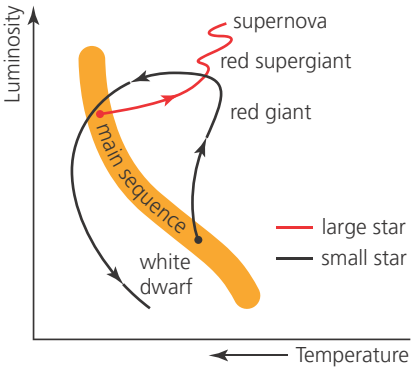


Figure 16.20

QUESTIONS TO CHECK UNDERSTANDING

33 Outline
- **a** why a main sequence star has a limited lifetime
- **b** what happens to cause it to evolve into a red giant.

34 The surface temperature of a main sequence star is 5600 K. Later it will evolve into a red giant. If its radius increases by a factor of 150 and it becomes 3000 times more luminous, estimate the red giant's surface temperature.

35 **a** What are the differences between red giants and red supergiants?
- **b** Why do these stars have limited lifetimes?

36 **a** Explain why the masses of all white dwarf stars are below a certain limit.
- **b** What is the name of that limit?

37 What is a neutron star?

38 Explain what the Oppenheimer–Volkoff limit has to do with the formation of black holes. Include neutron degeneracy pressure in your explanation.

NATURE OF SCIENCE

Evidence

Astronomy is unique among sciences in that it is mostly concerned with making observations and collecting evidence, rather than designing experiments in which variables are altered or controlled. Observation of the spectra from stars leads to conclusions about their composition and temperatures, as well as enabling astronomers to determine their speeds and develop theories about the birth and evolution of the whole universe.

16.3 Cosmology

Revised

Essential idea: 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 (cosmos).

Redshift (*z*)

Revised

- When the line spectra detected from distant galaxies (or stars) are compared to the line spectra from the same elements emitted on Earth, all the observed wavelengths (and frequencies) are slightly different. Figure 16.21 shows both possibilities: compared to the central spectrum (from a nearby source), the lines in the top spectrum are shifted to longer wavelengths, and in the lower spectrum the lines are shifted to shorter wavelengths.
- In most cases, the change in wavelength, $\Delta\lambda$, is an increase.
- Because red is at the higher wavelength end of the visible spectrum, this effect is commonly known as a **redshift**. The effect is not limited to visible light.
- There are also a few examples of observed wavelengths being decreased. That effect is called a **blueshift**.
- The amount of redshift, z, is defined by the equation $z = \frac{\Delta\lambda}{\lambda_0} = \frac{\lambda - \lambda_0}{\lambda_0}$, where λ_0 is the initial wavelength at the source and λ is the observed wavelength.

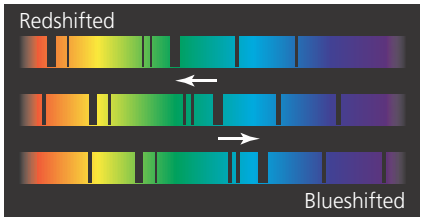


Figure 16.21

- For example, if the 656 nm line on the hydrogen spectrum was detected from a distant galaxy with a wavelength of 672 nm, the redshift would be $z = \frac{16}{656} = 0.024$ (a ratio, so it has no unit).
- Redshift has similarities to the Doppler effect (Chapter 9, Section 9.5 for HL students), in which the wavelength of a source of sound that is moving away from us is increased. Figure 16.22 shows an example, but there are important differences, which will be explained later.

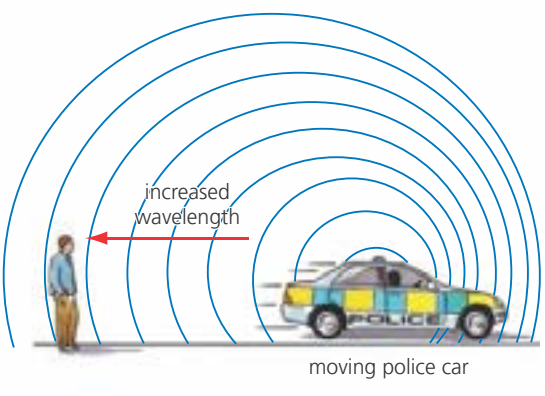


Figure 16.22

Key concept

The wavelengths of the lines in the spectra observed from distant galaxies and stars are (usually) slightly greater than those observed from sources on Earth. This is called a *redshift*.

The amount of redshift, z, is defined to be $z = \frac{\Delta\lambda}{\lambda_0}$.

The Big Bang model

Revised

- Redshift is evidence that distant galaxies and Earth are moving apart: **the expansion of the universe**.
- The magnitude of a redshift can be shown to be approximately equal to the ratio of the **recession speed**, v, to the speed of light $z = \frac{\Delta\lambda}{\lambda_0} \approx \frac{v}{c}$. (This equation was used in Chapter 9, Section 9.4.)
- Returning to the previous example, a redshift of $z = 0.024$ can now be seen to be characteristic of a galaxy with a recession speed of $v = 0.024c$ or 7.2×10^6 ms^{-1}.
- Very fast moving, distant galaxies have redshifts that are greater than one, but the highlighted equation involving v is an approximation that is only valid for $v << c$, so it cannot be used to determine the recession speed of such galaxies. (Other equations are possible, but they are not included in this course.)
- The light from a small number of stars and galaxies is blueshifted because their rotational speed within their galaxy or cluster of galaxies is greater than the recession speed of the whole system.

Key concept

If a shift is to a longer wavelength (a redshift), we know that the distance between the galaxy and Earth is increasing. We say that the galaxy is *receding* from Earth.

Recession speed, v, and the amount of redshift are linked by the equation $z = \frac{\Delta\lambda}{\lambda_0} \approx \frac{v}{c}$.

Key concept

When the light from a large number of galaxies is studied, we find that they nearly all have redshifts and so are receding from Earth. This can only mean that the universe is expanding.

Describing both space and time as originating with the Big Bang

- The distance of galaxies from Earth can be determined by using Cepheid variables as 'standard candles' (as explained in Section 16.2). The recession speeds can then be compared to the distances (see Figure 16.23).
- We can see from Figure 16.23 that the recession speed of a galaxy is proportional to its distance from Earth.
- Moving backwards in time, the conclusion must be that all the galaxies were together at the same place at some point in time. This is the (hot) *Big Bang model of the universe*.
- The galaxies are not moving from a central position into a pre-existing space, rather space itself is expanding.
- Before the Big Bang model of the universe became fully accepted, it was widely believed that the universe was uniform and static (and infinite). This is sometimes called the **Newtonian model of the universe.**

Key concepts

The **Big Bang model of the universe**: The universe began at one point at a particular time. It was incredibly hot and has been expanding and cooling ever since.

It is important to realise that the Big Bang was the creation of everything, including both space and time.

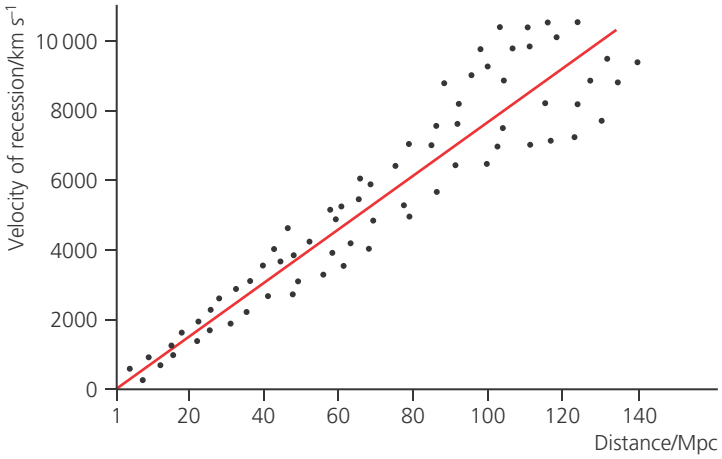


Figure 16.23

QUESTIONS TO CHECK UNDERSTANDING

39 A spectral line of oxygen has a wavelength of 5.38×10^{-7} m. When detected on Earth, the wavelength has changed by 0.11×10^{-7} m.
- **a** What was the value of the wavelength detected?
- **b** Determine the redshift of the galaxy.
- **c** Calculate the recession speed of the galaxy.

40 A distant galaxy is moving away from Earth with a speed of 5 300 km s^{-1}.
- **a** What redshift will be observed in radiation from this galaxy?
- **b** If a spectral line of wavelength 442 nm was observed on Earth from this galaxy, with what wavelength was it emitted?

41 **a** Explain how the expansion of space results in a redshift of the radiation received from distant galaxies.
- **b** Distinguish between this kind of redshift and the Doppler effects that can be detected from sources on Earth.

42 Explain, with the help of a diagram, how it is possible for the radiation from a few galaxies, or stars, to be blueshifted.

43 Explain how the study of redshifts provides evidence that
- **a** the universe is expanding
- **b** the universe began at one point at a particular time.

Key concept

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.

Key concept

The Big Bang Theory provides us with the correct interpretation of redshift: received wavelengths are greater than emitted wavelengths because space has expanded during the time of the radiation's journey. (This kind of redshift is known as *cosmological redshift*.)

Common mistake

The universe has no centre and no visible edge. The expansion of space means that all objects are moving apart from each other. The Earth is not in a special position, redshift observations would lead to the same conclusion no matter where they were made.

Expert tip

Cosmological redshift should not be confused with the true Doppler effect: a source and observer moving further apart (or closer together) in unchanging space (as in Figure 16.22).

Hubble's law

Revised

- The relationship shown by the straight line on the graph in Figure 16.23 is known as Hubble's law.
- The value of the Hubble 'constant' will not change over the course of, for example, a human lifetime, but we cannot assume that it always had, and will always have, the same value. (The subscript 0 is used to signify that it is the value of H at the present time.)
- Using Figure 16.23, $H_0 \approx \frac{9000}{120} = 75\,\text{km s}^{-1}\,\text{Mpc}^{-1}$. The latest value (2016) determined from measurements from the Hubble space telescope has $H_0 = 73 \pm 2\,\text{km s}^{-1}\,\text{Mpc}^{-1}$.
- There is significant uncertainty in the value of the Hubble constant because of uncertainties in determinations of distance, but also because of the movement of the observed galaxies within their clusters.
- Hubble's law enables the distances to galaxies to be estimated from measurements of the redshifts in the radiation received from them.

Key concept

Hubble's law: the current velocity of recession, v, of a galaxy is proportional to its distance from Earth, d.

As an equation: $v = H_0 d$ where H_0 is known as the (current value of the) **Hubble constant** (the gradient of the graph).

Estimating the age of the universe by assuming a constant expansion rate

- Hubble's law can be used to estimate the age of the universe: $T = \frac{d}{v}$, leads to $T \approx \frac{1}{H_0}$, although this calculation is an approximation because it assumes that the universe has always been expanding at exactly the same rate (constant H_0).
- Using $H_0 = 73\,\text{km s}^{-1}\,\text{Mpc}^{-1}$ leads to an estimated age of the universe, $T = 4.3 \times 10^{17}\,\text{s}$ (or 1.4×10^{10} years).
- The universe is bigger than we are able to observe. This is because we are limited by the speed of light. The distance to the 'edge' of the **observable universe** equals the age of the universe multiplied by the speed of light: $4.3 \times 10^{17} \times 3.0 \times 10^{8} \approx 10^{26}\,\text{m}$, but this figure needs to be increased to allow for the expansion of space since the Big Bang. The accepted value $\approx 4.4 \times 10^{26}\,\text{m}$.

Cosmic microwave background (CMB) radiation

Revised

- Astronomers were able to use data concerning the Big Bang expansion of the universe to predict its average temperature to be 2.76 K.
- We know that all objects emit radiation which is characteristic of their temperature. Using Wien's law ($\lambda_{max} T = 2.9 \times 10^{-3}\,\text{m K}$), we can determine the value of the wavelength at which radiation intensity is maximized for a temperature of 2.76 K: $\lambda_{max} = 1.1 \times 10^{-3}\,\text{m}$, which is in the microwave section of the electromagnetic spectrum.

Describing the characteristics of CMB radiation

- Figure 16.24 shows the spectrum at 2.76 K from any object which approximates to a black body. This radiation was found to be arriving at Earth (and by implication, anywhere else) equally from all directions (**isotropic**).
- Tiny variations were discovered many years later and are considered to be of great importance. (See Section 16.5 (HL).)

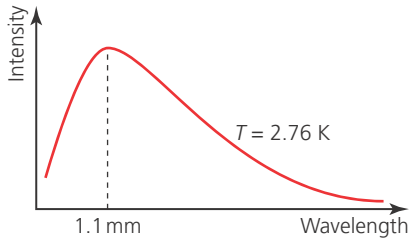


Figure 16.24

Explaining how the CMB radiation is evidence for the Hot Big Bang

- Alternatively, the current wavelength of CMB can be considered as a consequence of the expansion of space, stretching the shorter wavelengths that were emitted billions of years ago.

Key concept

Isotropic CMB radiation confirms that the average temperature of the universe is 2.76 K, as predicted by the Big Bang Theory.

QUESTIONS TO CHECK UNDERSTANDING

44 Referring back to Question 39, estimate the distance of the galaxy from Earth.

45 Convert $H_0 = 73\,\text{km s}^{-1}\,\text{Mpc}^{-1}$ to SI units.

46 Give two reasons why using the equation $T = \frac{1}{H_0}$ to determine the age of the universe will only produce an approximate answer.

47 The currently accepted age of the universe is 13.8 billion years. What is the corresponding value of the Hubble constant?

48 A very long time ago, the average temperature of the universe was 5 K. What was the value of the wavelength that had peak intensity at that time?

49 Explain why the discovery of isotropic microwave background radiation was considered to be very important.

The cosmic scale factor (*R*)

Revised

- Figure 16.25 is a two-dimensional representation of the same galaxies at three different times in an expanding universe. The pattern remains the same but the *scale* changes because of the expansion of space (comparable to changing the scale on a GPS map).
- The 'before' drawing has a scale which is 0.55 × the scale of the 'now' drawing, and the 'later' drawing has a scale which is 1.45 × the scale of 'now'.
- Astronomers use a *cosmic scale factor, R*, to describe the changing dimensions of the universe.
- If the cosmic scale factor now is assumed to be 1, at a time when the separation of any two galaxies was half its current value, then $R = 0.5$ at that time, etc.

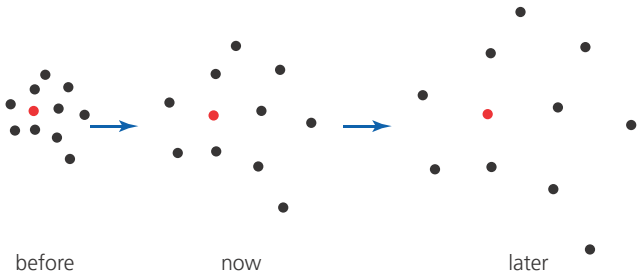


Figure 16.25

- Clearly, the cosmic scale factor changes with time, and a graph of cosmic scale factor–time is a convenient way of representing the expansion of the universe (see Figure 16.26, page 129).
- Consider radiation emitted with a wavelength λ_0 from a galaxy millions of years ago when the cosmic scale factor was R_0. By the time that the radiation is received on Earth (now), the wavelength and cosmic scale factor have increased to λ and R.
- Since $\frac{R}{R_0} = \frac{\lambda}{\lambda_0}$, and $z = \frac{\lambda - \lambda_0}{\lambda_0} = \frac{\lambda}{\lambda_0} - 1$; redshift, $z = \frac{R}{R_0} - 1$.
- For example, if the cosmic scale factor was 0.45 at a time 6.0×10^{10} years ago, the redshift between then and now, $z = 1.22$, so that radiation emitted with a wavelength of 588 nm would be shifted by 719 nm (to a new value of 1307 nm) by the time it was received on Earth 6.0×10^{10} years later (now).

Key concept

The **cosmic scale factor** is a convenient way of representing the expansion of the universe.

The cosmic scale factor (at a time *t*),

$$R = \frac{\textit{separation of two points at time } t}{\textit{separation of the same two points now}}$$

(This is a ratio, so there is no unit.)

Variations in the cosmic scale factor are closely connected to the numerical value of redshift:

$$z = \frac{R}{R_0} - 1.$$

Solving problems involving *z*, *R* and Hubble's law

QUESTIONS TO CHECK UNDERSTANDING

50 Radiation was emitted from a distant galaxy with a wavelength of 6.87×10^{-7} m and detected on Earth with a wavelength of 7.29×10^{-7} m.

a Calculate the redshift involved.

b What was the cosmic scale factor at the time the radiation was emitted?

51 The radiation from a very distant galaxy has a redshift of 6.1.

a What was the cosmic scale factor at the time the radiation was emitted?

b Estimate the size of the observable universe at that time (current value $\approx 4.4 \times 10^{26}$ m).

52 A galaxy is located 500 Mpc from Earth.

a Determine the redshift in radiation from this galaxy.

b What was the cosmic scale factor at the time the radiation was emitted?

History and future of the universe

Revised

- As we have discussed, astronomers have convincing evidence that the universe started with a 'Big Bang' and continues to expand, but the pattern and timing of that expansion, and how it may continue in the future, are still the subject of considerable research.
- To begin an analysis, we may assume that the expansion of the universe is opposed only by the force of gravity, and that the rate of expansion and the time for which it will continue depend on the size of the gravitational forces, which in turn depend on the total mass in the universe.
- Based on the idea that the kinetic energy of galaxies is being transferred to gravitational potential energy as the universe expands, Figure 16.26 shows some general possibilities for the evolution of the universe.
  - The orange line represents a universe that would reach a maximum size and then contract.
  - The green line represents a universe that would expand for ever, but at a rate that reduces to zero after infinite time.
  - The blue line represents a universe that would continue to expand for ever (but at a decreasing rate).
- The red line in Figure 16.26 represents an *accelerating universe* in which the rate of expansion is increasing. Such an expansion cannot be explained only in terms of gravitational forces (see below).
- Ideas about the expansion of the universe are discussed in more detail for HL students in Section 16.5.

Key concept

The future of the universe depends on the total mass that it contains. By considering a simple model in which kinetic energy is transferred to gravitational potential energy, it is possible to identify several different possibilities.

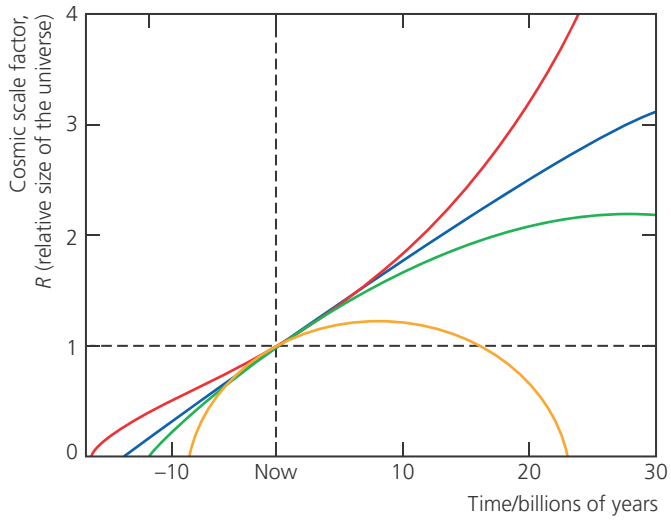


Figure 16.26

The accelerating universe

- In recent years, astronomers have collected convincing evidence that the universe is in fact 'accelerating'. This evidence comes from the observation of distant type Ia supernovae, as seen in Figure 16.27.
- The luminosities of all type Ia supernovae are known to be the same (explained in Section 16.4), so that they can be used as 'standard candles' and their distances from Earth can be calculated from a measurement of their apparent brightness.
- However, calculations using Hubble's law predict smaller distances. In other words, the universe is expanding quicker than previously believed.

Figure 16.27

- 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.

QUESTIONS TO CHECK UNDERSTANDING

53 Explain why the future of the universe depends on the amount of mass it contains.

54 Summarize why observations made on distant supernovae forced astronomers to change their theories about the evolution of the universe.

55 Explain

a the term 'accelerating universe'

b why a concept like dark energy was needed to explain it.

Key concepts

Type Ia supernovae always have the same known luminosity, so that they can be used to determine the distance to the galaxies in which they occur $\left(\text{using } b = \frac{L}{4\pi d^2}\right)$.

Calculations using data from type Ia supernovae confirm that distant galaxies are further away than expected, and even further away than expansion of the universe at a constant rate would predict.

The existence of **dark energy** has been proposed as an explanation of the accelerating universe.

NATURE OF SCIENCE

Occam's razor

The principle of *Occam's razor* has been interpreted in many different ways. Its essence is that simplicity is best. If there are two or more theories that explain the same observed facts, then the simplest one is the better choice (until and unless further evidence contradicts it). The Big Bang model began as a simple model (simpler than others at the time), although it does not explain the moment of creation and it has also been complicated by more recent discoveries.

16.4 Stellar processes (higher level only)

Revised

Essential idea: The laws of nuclear physics applied to nuclear fusion processes inside stars determine the production of all elements up to iron.

The birth of stars

- From Section 16.1, we know that a star is born in part of a nebula where gravity has pulled *interstellar matter* closer together and the gas atoms have gained the very high kinetic energies needed for nuclear fusion to occur. We will now consider this in more detail.
- The slow inwards collapse of clouds of interstellar matter (because of gravitational forces) is opposed by the random motions of the particles, creating an outwards gas pressure. In order for star formation to begin, the total mass of the gas cloud has to be great enough to create sufficient inwards gravitational forces to overcome the gas pressure. Then part of the cloud will collapse inwards until nuclear fusion begins and opposes the collapse with greater thermal gas pressure and radiation pressure outwards.
- Interstellar matter is not totally uniform (homogeneous) and conditions for star formation will be more favourable where the density is greatest. Conditions will be affected by neighbouring stars or the shock waves from supernovae.
- Figure 16.28 shows a 'star nursery' called the Elephant's trunk nebula.

Figure 16.28

The Jeans criterion

- For a given temperature, the *minimum* mass required of a cloud of interstellar matter for star formation is called the **Jeans mass**, M_J.
- The *Jeans criterion*: the collapse of an interstellar cloud to form stars can only begin if its mass $M > M_J$.

Key concept

In order for a star to be formed from an interstellar gas cloud, the mass of the cloud must be greater than a certain value, known as the *Jeans mass*.

The Jeans mass is temperature dependent.

- The Jeans mass is large enough for the formation of many stars from the same nebula.
- The value of the Jeans mass can be determined from an appreciation that collapse begins if the magnitude of the gravitational potential energy of all the mass involved is greater than the kinetic energy of the particles. However, a derivation of the Jeans mass is not required for the IB Physics course.
- The Jeans mass is dependent on the temperature, T, and particle density, n, and for a cloud consisting of hydrogen it can be estimated from: $M_J \approx (3 \times 10^4)\sqrt{\left(\frac{T^3}{n}\right)}$. The calculated mass is in multiples of the solar mass (2.0×10^{30} kg). This equation does not need to be remembered.

Common mistake

It is important to realise that the Jeans mass is very temperature dependent: a greater mass is required for star formation at higher temperatures.

Applying the Jeans criterion to star formation

- The conditions for star formation are more favourable if the temperature is lower and the particle density is greater. This corresponds to a smaller Jeans mass.
- As an example, using the last equation, for a temperature of 100 K and a hydrogen cloud of density 10^{10} atoms per cubic metre, the Jeans mass is approximately 300 solar masses; for a temperature of 50 K and the same density it would be easier for a star to form, which is shown by a reduced Jeans mass of about 100 solar masses.

QUESTIONS TO CHECK UNDERSTANDING

56 Explain why gas pressures increase at greater temperatures and densities.

57 Why is the Jeans mass smaller for a lower temperature (at the same density)?

58 Estimate the number of molecules of hydrogen in one cubic metre at atmospheric pressure and room temperature on Earth.

59 A hydrogen cloud had a particle density of 10^{12} atoms per cubic metre. Use the equation $M_J \approx (3 \times 10^4)\sqrt{\left(\frac{T^3}{n}\right)}$ to compare the Jeans masses for this cloud with another which has a particle density of 10^9 atoms per cubic metre at the same temperature.

Expert tip

Continuing the previous example, a Jeans mass of 300 solar masses ($= 6 \times 10^{32}$ kg) applies to a hydrogen cloud of temperature 100 K and density 10^{10} atoms per cubic metre ($= 1.7 \times 10^{-17}$ kg m^{-3}). This corresponds to a volume of about 4×10^{49} m^3, which has a diameter of 4×10^{16} m ≈ 4 ly (assuming spherical shape).

Main sequence lifetimes

Revised

- A typical main sequence star begins its lifetime with about 75% hydrogen, which is spread throughout the star. Most of the rest of the star is helium, but there are traces of heavier elements left over from supernovae. The star is hottest at its *core* and that is where the hydrogen nuclei have enough kinetic energy to fuse into helium (see Figure 16.29).
- Helium is denser than hydrogen so it collects at the centre of the star. The amount of hydrogen in the core reduces and eventually it is not enough to sustain the nuclear fusion. This is the beginning of the end of the star's life on the main sequence.
- The core collapses inwards and this results in significantly increased temperatures as gravitational potential energy is transferred to kinetic energy of particles. This heats up a 'shell' around the core enough for a considerable increase in the fusion of hydrogen into helium in that shell.
- The result is a relatively quick increase in the rate of energy transformation so that the star undergoes significant increase in size and its outer layers cool: it becomes a red giant (or red supergiant).

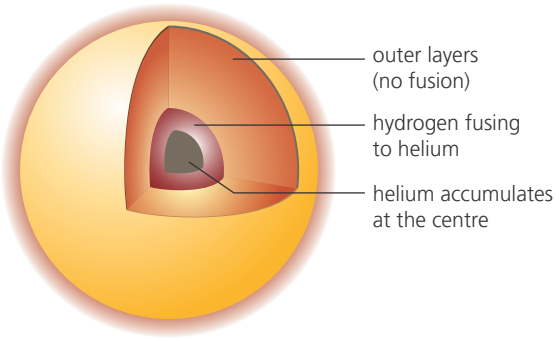


Figure 16.29

Applying the mass–luminosity relation to compare lifetimes on the main sequence relative to that of our Sun

- In Section 16.2, we noted that the luminosity of a star is related to its mass by $L \propto M^{3.5}$, showing that more massive stars are *much* more luminous. This is because the temperatures and rate of fusion in a more massive star are much greater since greater gravitational forces accelerate particles to higher speeds.
- If we assume that the total energy available from nuclear fusion is proportional to the total mass of a star, then its average luminosity, L (power $= \frac{\text{energy}}{\text{time}}$) $\propto \frac{M}{T}$, where T is its lifetime as a main sequence star.
- So that $M^{3.5} \propto \frac{M}{T}$, leading to $T \propto \frac{1}{M^{2.5}}$ or $TM^{2.5} = \text{constant}$. 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).
- For example, using the last equation, we can show that a star with twice the mass of the Sun will have a lifetime about six times shorter.

Key concept

Main sequence stars of greater mass have much greater luminosities ($L \propto M^{3.5}$ from Section 16.2). Assuming that $L \propto \frac{M}{T}$, where T is the main sequence lifetime, we can see that $T \propto \frac{1}{M^{2.5}}$.

QUESTIONS TO CHECK UNDERSTANDING

60 Summarize why more massive main sequence stars have relatively shorter lifetimes.

61 The expected lifetime of the Sun is about 1×10^{10} years. Estimate the lifetime of a star which has a mass ten times smaller than the Sun.

62 a Determine the relative mass of a star which will have a lifetime which is only 0.01% of that of the Sun.

b Use the HR diagram to estimate the surface temperature of the star.

Nuclear fusion

Revised

Fusion in main sequence stars

- Nuclear fusion in main sequence stars is predominantly hydrogen into helium, but it is not a simple one-step process, as is explained below. Fusion involves the release of a large amount of energy in the form of the kinetic energy of the nuclei, gamma rays and neutrinos.
- In Section 16.1, it was summarised as $4\,{}^1_1\text{H} \rightarrow {}^4_2\text{He} + 2\,{}^0_1\text{e}$ + neutrinos and photons, with the release of 27 MeV of energy, but now further details can be provided. This is commonly called the *proton–proton cycle*.
  - Two protons fuse to make a ${}^2_1\text{H}$ (deuterium) nucleus. In this process, a positron and an (electron) neutrino are emitted:
    ${}^1_1\text{H} + {}^1_1\text{H} \rightarrow {}^2_1\text{H} + {}^0_1\text{e}^+ + {}^0_0\nu_e$.

Key concept

The fusion of hydrogen to helium in *smaller main sequence stars* (like the Sun) is a three-stage process known as the **proton–proton cycle**.

For *larger main sequence stars*, which have higher core temperatures, a different process is involved. It is known as the CNO cycle because atoms of carbon, nitrogen and oxygen are involved.

- Then, the deuterium nucleus fuses with another proton to make He-3. In this process, a gamma ray photon is emitted. ${}^{2}_{1}\text{H} + {}^{1}_{1}\text{H} \rightarrow {}^{3}_{2}\text{He} + {}^{0}_{0}\gamma$.
- Finally, two He-3 nuclei combine to make He-4. Two protons are released in this reaction. ${}^{3}_{2}\text{He} + {}^{3}_{2}\text{He} \rightarrow {}^{4}_{2}\text{He} + 2\,{}^{1}_{1}\text{H}$
- Figure 16.30 shows the combined process.

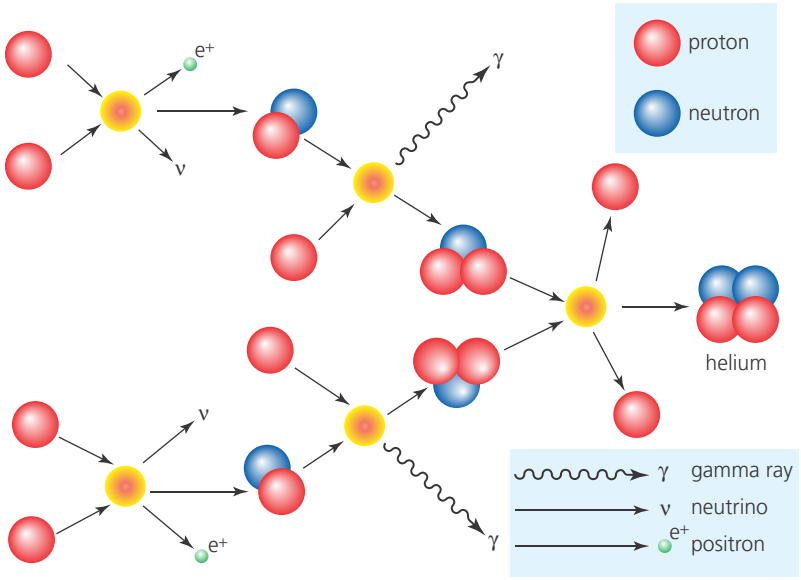


Figure 16.30 The proton–proton cycle.

Nucleosynthesis off the main sequence

- **Nucleosynthesis** is the term used to describe the creation of larger nuclei of different elements from smaller nuclei and nucleons.

Describing the different types of nuclear fusion reactions taking place off the main sequence formation

- Red giants with larger masses have cores with higher temperatures, so that the nuclei have greater kinetic energy and can overcome the greater repulsive forces that act between nuclei of greater charge.
  - For stellar masses less than $4M_{\odot}$ (red giants), the core temperature can reach over 6×10^{8} K and this is large enough for the nucleosynthesis of carbon and then oxygen. For example, ${}^{4}_{2}\text{He} + {}^{12}_{6}\text{C} \rightarrow {}^{16}_{8}\text{O}$. Helium will still be produced in an outer layer.
  - 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. Outside of the core, there will be layers rich in oxygen, carbon, helium and hydrogen. Such stars will end their lives as white dwarfs (see Section 16.2).
  - 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. Such stars will end their lives as neutron stars or black holes (see Section 16.2).
- The structure of stars off the main sequence is layered as the more massive nuclei are found closer to the centre. A red supergiant will have the most layers, as shown in Figure 16.31.
- All these nuclear fusion processes result in the *emission* of energy because the synthesised nuclei have increased binding energies per nucleon (see Chapter 7, Section 7.2). However, binding energy per nucleon has its maximum values for iron and nickel, which means that more massive nuclei (than iron) cannot normally be produced by nuclear fusion and they cannot provide the fuel source for stars.

Key concept

The elements in the universe have all been created from lighter nuclei in nuclear reactions at very high temperatures. This process is called nucleosynthesis.

Key concepts

The temperatures in the cores of red giants or supergiants are sufficient to cause the fusion of some heavier elements.

Elements heavier than iron cannot be produced in this way.

More massive nuclei will move towards the centre of the star, resulting in spherical layers around the core in which different nuclei are concentrated.

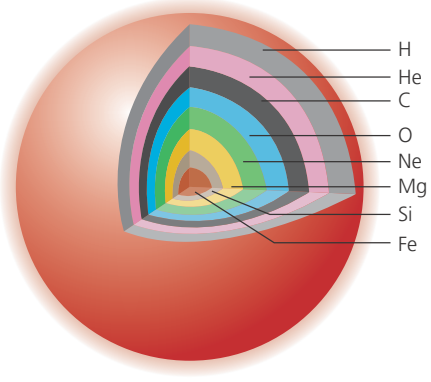


Figure 16.31

QUESTIONS TO CHECK UNDERSTANDING

63 What is the predominant nuclear fusion process taking place in main sequence stars?

64 a Estimate the loss of mass from the Sun every year assuming its luminosity is 3.4×10^{26} W.

b Estimate the number of helium atoms fused from hydrogen–1 every second if each fusion releases 27 MeV of energy.

65 Explain

a why larger red giants have greater temperatures in their cores, and

b why a star of mass $8M_{\odot}$ cannot produce nuclei more massive than magnesium.

66 a Write a nuclear equation to represent the nuclear fusion of oxygen-16 with helium-4.

b Explain why this fusion releases energy.

67 Use the concept of binding energy per nucleon to explain why a nuclide like $^{64}_{30}\text{Zn}$ cannot be synthesized by fusion from less massive nuclei.

Describing the formation of elements in stars that are heavier than iron including the required increases in temperature

- The formation of elements heavier than iron involves the processes of *neutron capture*.
- Many neutrons are released during fusion processes in stars. An example is $^{22}_{10}\text{Ne} + {}^{4}_{2}\text{He} \rightarrow {}^{25}_{12}\text{Mg} + {}^{1}_{0}\text{n}$.
- Neutron capture is possible because neutrons are uncharged and are therefore unaffected by electric forces within an atom. At very high temperatures neutrons have enough kinetic energy to get very close to a nuclei and then they can be attracted by the strong nuclear force.
- When one (or more) neutrons is captured by a nucleus, the atom becomes a more massive isotope of the original element. The new nucleus will probably be unstable and liable to decay by beta-negative emission (Chapter 7) to a different element with a *greater* proton number.
- The nucleosynthesis of lighter elements may be by fusion or neutron capture, but neutron capture is the *only* way in which nuclei more massive than iron can be synthesised.
- Neutron capture example: $^{114}_{48}\text{Cd} + {}^{1}_{0}\text{n} \rightarrow {}^{115}_{48}\text{Cd}$ + gamma ray; then $^{115}_{48}\text{Cd} \rightarrow {}^{115}_{49}\text{In} + {}^{0}_{-1}\text{e} + \bar{\nu}_e$. (This is an example of *s-process* neutron capture, as explained below.)

Key concept

Neutron capture is a process in which one or more neutrons is absorbed by a nucleus of an atom.

Expert tip

The number of neutrons passing through unit area every second is often called the *neutron flux*. Neutron capture is obviously more likely with a greater neutron flux. The possibility of neutron capture is often represented by the term *neutron capture cross-section*. A larger cross-section (for a certain neutron speed) implies a greater probability of capture.

Qualitatively describe the s and r processes for neutron capture

- Clearly nucleosynthesis by neutron capture is a two-step process: firstly neutron capture, then beta negative decay. However, these two processes can occur at very different rates.
- **Slow neutron capture (s-process)** occurs at a relatively low neutron density (flux) and at an intermediate stellar temperature. These conditions are found in red giants. After a neutron has been captured, the new nucleus will typically have plenty of time (maybe hundreds of years) to undergo beta decay before further neutron capture might occur. The nucleosynthesis of In-115, described by the equation above, is an example of the s-process. The process is also represented in Figure 16.32.

Key concept

We can identify two extremes of neutron capture: (1) the nuclei decay at a rate much quicker than the rate of neutron capture (*s-process*); (2) the nuclei decay much slower than neutrons are captured (*r-process*).

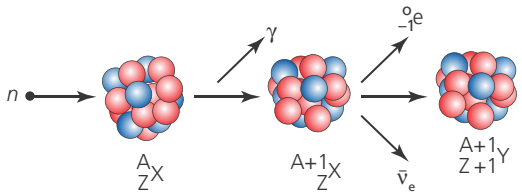


Figure 16.32

- **Rapid neutron capture (r-process)** occurs at extreme temperatures and neutron densities (flux). These conditions are only found in supernovae (see next section). Many neutrons are captured by the same nucleus before there is enough time for beta decay to occur. This process occurs very quickly, maybe only in minutes or less. The heaviest elements in the universe are formed this way. As an example, iron-56 may capture five neutrons one-by-one until it has become iron-61, but this particular nuclide is much less stable than the lighter ones and decays to cobalt-61.
- Figure 16.33 shows how an example of the r-process appears on a part of a chart of the nuclides. A nucleus of ytterbium-188 captures eight neutrons until it becomes ytterbium-196, which then decays (by beta negative decay) to lutetium-196.

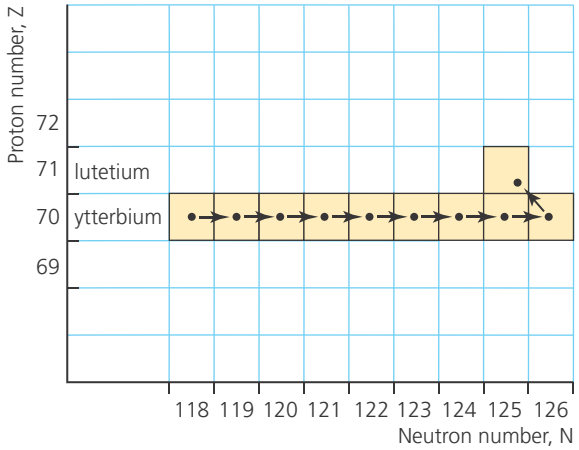


Figure 16.33

QUESTIONS TO CHECK UNDERSTANDING

68 Distinguish between slow neutron capture and rapid neutron capture.

69 Write an equation to represent the following s-process: the nucleosynthesis of germanium-72 from $^{71}_{31}Ga$.

70 a Explain how the nuclide $^{209}_{83}Bi$ might be formed from $^{204}_{82}Pb$ by the r-process.

b Why are extremely high temperatures needed for the r-process?

c Where do these temperatures occur?

Type Ia and II supernovae

Revised

- Supernovae were described in Section 16.2 as the events at the end of the lives of red supergiants that result in the formation of neutron stars or black holes (these are known as *type II* supernovae). Another kind of supernova was mentioned in Section 16.3 in connection with the accelerating expansion of the universe: *type Ia* supernovae .

Distinguishing between type Ia and II supernovae

- If the gravitational attraction from a white dwarf star is great enough to attract a large amount of mass from another nearby star, its mass may increase sufficiently that electron degeneracy pressure (see Section 16.2) is no longer sufficient to resist its collapse. Figure 16.34 shows an artist's impression of this process.
- As a result of the collapse, there is sudden and considerable nuclear fusion, resulting in the type Ia supernova. This event always occurs when the mass of the white dwarf reaches a precise value (the *Chandrasekhar limit*), so that the resulting luminosity is always the same (about 10^{10} times greater than the Sun). This is why type Ia supernovae are used as *standard candles* for determining the distance to remote galaxies.
- *Type II supernovae* occur when nuclear fusion comes to an end in a red supergiant star.
- The temperature may rise to 10^{11} K and the nuclei in the core can get deconstructed back to protons, neutrons, electrons, photons and 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. As described above, rapid neutron capture occurs, resulting in the creation of the heavier elements. The remaining core will become a neutron star or a black hole.
- Different types of supernovae may be distinguished from each other by how their luminosity varies with time. Figure 16.35 shows examples of their 'light curves'. Note that the luminosity scale is logarithmic.

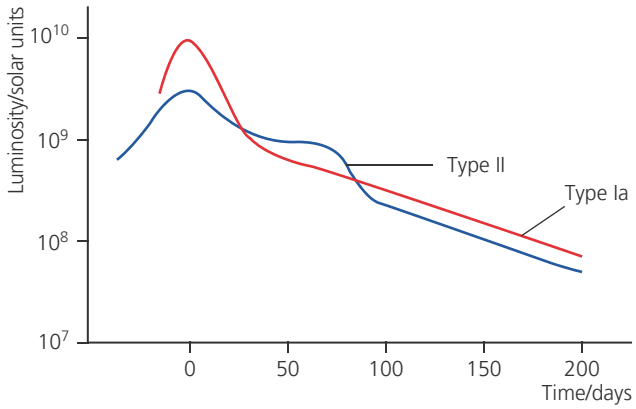


Figure 16.35

Key concepts

A **supernova** is a sudden, enormous and relatively short-lived explosion of a star.

Type Ia supernovae occur when a *white dwarf* star attracts enough matter from a smaller nearby star (in a binary system) to make it collapse.

A **type II supernova** is the result of a sudden inwards collapse of the core of a red supergiant when the fusion processes stop.

Figure 16.34

QUESTIONS TO CHECK UNDERSTANDING

71 **a** What is a white dwarf star?

b Under what condition can a white dwarf star become a supernova?

72 Explain why all type Ia supernovae have the same luminosities.

73 Outline the cause of a type II supernova.

74 Estimate the maximum apparent brightness of a type Ia supernova which occurs at a distance of 500 Mpc from Earth. (Luminosity of the Sun is 3.8×10^{26} W.)

75 Use Figure 16.35 to estimate

a how much greater the maximum luminosity of a type Ia supernova is compared with a type II

b how long it takes for the luminosity of a type Ia supernova to drop to 1% of its maximum value.

NATURE OF SCIENCE

Observation and deduction

Astronomers have direct evidence (observation of spectra) of the existence of the same chemical elements in the stars and interstellar space, as can be found on Earth. How those elements were created was deduced from an understanding of nuclear processes on Earth.

16.5 Further cosmology (higher level only)

Revised

Essential idea: 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.

The cosmological principle

Revised

- Most of this option involves descriptions of the different objects that can be observed from Earth. However, perhaps the most basic principle in cosmology is that the universe, if it is viewed on a large enough scale (hundreds of millions of light years), is essentially the same everywhere.

Describing the cosmological principle and its role in models of the universe

- The cosmological principle describes a universe that is *homogeneous* and *isotropic* (on the large scale).
- **Homogeneous** means uniform, the same everywhere. Observers in different locations will make the same observations.
- **Isotropic** means the observations made in different directions from the same location are the same.
- An isotropic universe must be homogeneous, but, in principle, a homogeneous universe does not have to be isotropic.
- The universe does not have an observable edge.
- The existence of, for example, superclusters of galaxies, may suggest that the cosmological principle has limitations, but most importantly, the principle implies that the laws of physics are the same everywhere and for all times, and that observations made *anywhere* in the universe would lead to the same conclusions.

Key concept

Astronomers believe that, when considered on the large scale, the universe is essentially the same everywhere. This is called the **cosmological principle**.

Key concepts

There is no 'special' location in the universe. The universe has no centre. There is no single location where the Big Bang occurred. The universe is *homogeneous*.

Observations made anywhere and in any direction are essentially the same. The universe is *isotropic* and has no edge.

The cosmological origin of redshift

Revised

There are two main reasons why radiation from a star or galaxy may be redshifted:

1. *Cosmological redshift* occurs because of the expansion of space, as discussed in Section 16.3.
2. The *Doppler effect* results in a redshift if a star is independently moving away from Earth, regardless of the expansion of space (but a blueshift if it is moving towards Earth).

When observing a star or galaxy, it is probable that cosmological redshift and the Doppler shift will both occur at the same time. For example, if a star is rotating in its galaxy towards Earth, while the galaxy as a whole is receding due to the expansion of space. The assessment of the rotational speed of stars within a galaxy is an important example of this (see next section).

Key concepts

Cosmological redshift: 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.

Doppler effect: The source of radiation and the observer are moving relative to each other (disregarding the expansion of space).

QUESTIONS TO CHECK UNDERSTANDING

76 a Why is it possible for us to see many more stars in the night sky in some directions, than in other directions?

b Does this fact contradict the cosmological principle?

77 Different parts of the universe, each about a hundred million light years (or more) in diameter, appear essentially the same.

a What term do we use to describe this observation?

b Estimate the percentage of the observable universe that each of those parts occupies.

78 Discuss why the cosmological principle suggests that the same laws of physics apply everywhere and for all time.

79 Explain why a homogeneous universe might not be isotropic.

80 a Why can there never be any cosmological blueshifts?

b Why are astronomers more likely to observe blueshifts from stars and galaxies which are relatively close to Earth?

Expert tip

There is another possible reason for a redshift. *Gravitational* redshifts occur as a result of radiation moving out of gravitational fields.

Mass and density of the universe

Revised

- Astronomers commonly refer to the *average density* of the universe.
- To estimate the density of the observable universe, astronomers can determine the masses and densities of individual galaxies in a sample region that is large enough that it can be considered homogeneous and representative of the entire universe.

Key concept

The mass of a galaxy may be estimated by adding the masses of the observed stars it contains, but that overlooks any mass that is not emitting radiation that we can detect. A better and quicker method is to use the mathematics of rotational dynamics.

Deriving rotational velocity from Newtonian gravitation

- The stars in a galaxy rotate around their common *centre of mass* and the rotational velocity (speed), v, of any star can be predicted from classical physics theory.
  - Consider Figure 16.36a. We saw in Chapter 10 that a mass, m, orbiting a larger spherical mass, M, at a distance r from its centre had a rotational (orbital) velocity, $v = \sqrt{\left(\frac{GM}{r}\right)}$, suggesting that rotational velocity may be approximately proportional to $\sqrt{\left(\frac{1}{r}\right)}$ for stars a long way from the centre of a galaxy. This was derived by equating the gravitational force to the force required for circular motion: $\left(\frac{GMm}{r^2}\right) = \frac{mv^2}{r}$. However, this equation can only be used for stars a long way from the concentration of mass near the centre of the galaxy.
  - For stars that are nearer to the centre of the galaxy, and surrounded by other stars, it is necessary to consider that as the distance from the centre increases, so too does the mass of stars within that distance. Consider Figure 16.36b, in which the shaded area represents a region close to the centre of a galaxy with uniform density ρ. M is the mass within a distance r from the centre and we can replace M in the previous highlighted equation with $\frac{4}{3}\pi r^3\rho$, which leads to $v = \sqrt{\left(\frac{4\pi G\rho}{3}\right)}r$ suggesting that rotational velocity may be approximately proportional to r for stars near the centre of a galaxy.
  - Combining these two results leads to the red theoretical line shown in Figure 16.37, which is known as a **rotational curve**.

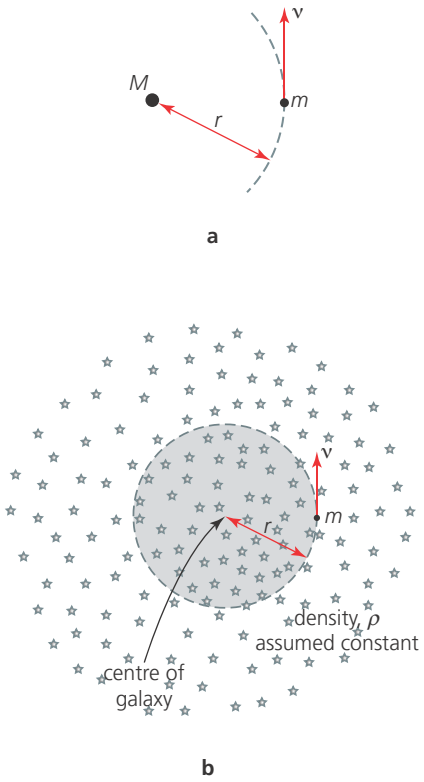


Figure 16.36

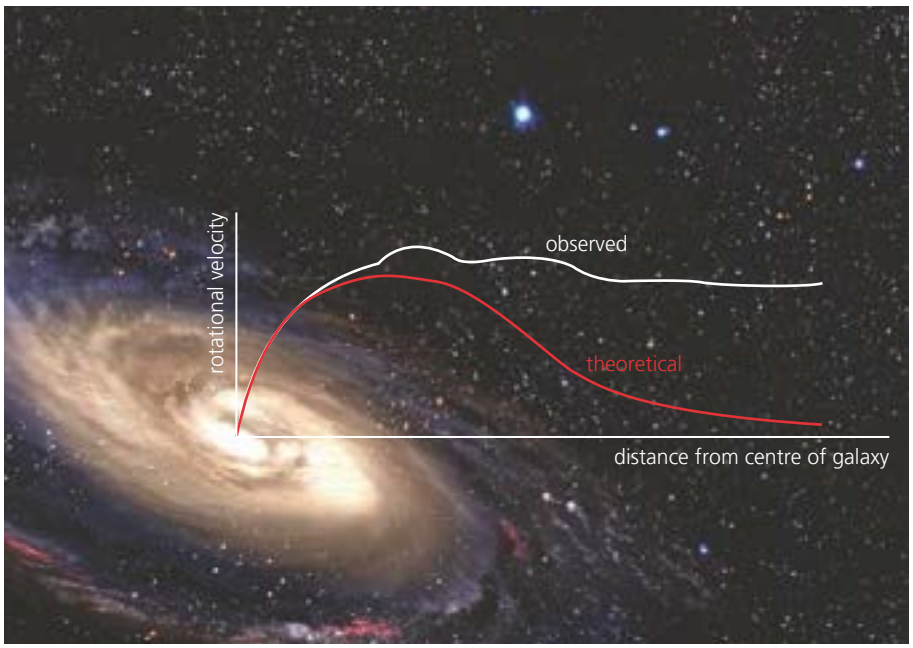


Figure 16.37

Key concept

An equation for the rotational velocity of a star *near the centre of a galaxy* may be derived by equating the gravitational force acting on it to the formula for centripetal force

$$v = \sqrt{\left(\frac{4\pi G\rho}{3}\right)}r$$

Rotation curves and the masses of galaxies

- The actual observed rotational speeds of the stars in a galaxy can be determined from Doppler shift measurements.
- The physics of rotational dynamics can be used with rotational speeds to determine the mass and average density of the galaxy.

Describing rotation curves as evidence for dark matter

- The theoretical rotation curve does not correspond to actual observations, shown in white in Figure 16.37. The outer stars in a galaxy have much greater rotational speeds than predicted from calculations involving the known masses in the galaxy. The explanation is that there must be a lot of mass in a galaxy which has not been detected. This is called *dark matter* and it is mostly found in the outer parts of the galaxy (in its 'halo').

Key concept

The outer stars in a galaxy rotate quicker than calculations involving the observed mass of the galaxy predict. This suggest that there is more mass in a galaxy than can be directly observed.

Dark matter

Revised

- Dark matter is believed to be about 85% of the total mass in the universe.
- Dark matter is a subject of much continuing research in astronomy as explanations for this 'missing' mass are sought. The theoretical possibilities are usually described under one of two categories:
  - **Weakly interacting massive particles** (WIMPs). These would be currently undiscovered particles created in enormous quantities at the time of the Big Bang. They would not be baryonic objects (not made from baryons: composed of quarks) and their only interaction would be via the nuclear weak force, making them extremely difficult to detect (like neutrinos). Describing them as 'massive' in this context means that the particles have mass, not that they are large.

Key concept

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.

Two types of explanation have been suggested: WIMPs and MACHOs.

- A less likely explanation for the enormous quantities of dark matter are MACHOs: **massive compact halo objects**. These would be composed of 'normal' baryonic matter located in galaxy halos. Possibilities include undetected neutron stars, dwarf stars and small black holes.

QUESTIONS TO CHECK UNDERSTANDING

81 a Derive the equation $v = \sqrt{\left(\frac{4\pi G\rho}{3}\right)}r$

b Explain why this equation cannot be used to predict the rotational speeds of stars near the edge of a galaxy.

82 Light from a star orbiting 512 ly from the centre of a galaxy has a maximum Doppler redshift, $z = 3.8 \times 10^{-4}$.

a Determine the rotational speed of the star.

b Assuming that its location is relatively close to the centre of the galaxy, estimate the average density of the centre of the galaxy.

83 The mass of the Milky Way galaxy is approximately 2.2×10^{42} kg and it radius is about 5×10^{20} m.

a Use these figures to estimate the average density of the Milky Way assuming (for simplicity) that it is spherical (in fact it is a spiral galaxy).

b Ignoring dark matter, estimate the rotational speeds of stars which are 5×10^{19} m from the centre of the galaxy (assume the average density close to the core is about ten times greater than the answers to (a)).

84 Explain the differences between WIMPs and MACHOs.

85 Why is most of the dark matter in a galaxy thought to be in its outer halo?

Expert tip

Another possibility for dark matter is the *axion*. This is a hypothetical low mass, uncharged particle proposed about 40 years ago, but which has still not been confirmed.

Expansion of the universe

Revised

- We saw in Section 16.2 that the rate of expansion of the universe depends on the total mass it contains. This is because the sizes of the gravitational forces opposing expansion depend on the masses involved.
- Of particular interest is the special situation in which the amount of mass known to be in the universe (including dark matter) predicts that it will expand forever, but the rate of expansion will reduce to zero after infinite time. This is described as a **flat universe**. (The term *flat* refers to the curvature of space, which is not included in this option).
- A flat universe can only occur if the amount of mass in the universe has an exact value.

Key concepts

An equation for the critical density, ρ_c, of the universe can be derived by equating the loss of kinetic energy of a mass to its gain in gravitational potential energy as it moves to infinity.

$$\rho_c = \frac{3H^2}{8\pi G}$$

Critical density

- The **critical density**, ρ_c, of the universe is the average density of matter which would result in a flat universe.

Deriving critical density from Newtonian gravitation

- Consider Figure 16.38 which shows a homogeneous spherical section of the universe of mass M, radius r and density ρ. 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.
- A theoretical value for the critical density can be obtained by considering that if the speed of the mass m reduces to zero at infinite time and distance, then its loss of kinetic energy equals the gain in gravitational potential energy:

$$\frac{1}{2}mv^2 = \frac{GMm}{r}.$$

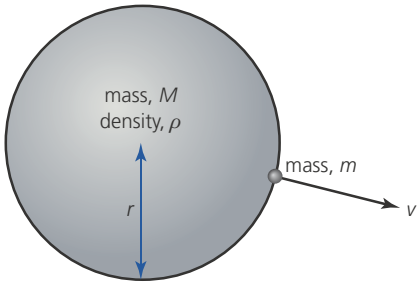


Figure 16.38

- Substituting $M = \left(\frac{4}{3}\right)\pi r^3 \rho_c$ and $v = Hr$, then rearranging, leads to $\rho_c = \frac{3H^2}{8\pi G}$.
- Using a value of $H = H_0 = 73\,\text{km s}^{-1}\text{Mpc}^{-1}$ leads to $\rho_c \approx 10^{-26}\,\text{kg m}^{-3}$, which is equivalent to about 6 atoms per cubic metre (assuming hydrogen is the dominant element present in the universe).

We live in a flat universe

- Astronomers have estimated the average density of the universe from observational data (as described above) and compared that value to theoretical value for the critical density. Perhaps surprisingly, we appear to live in a universe which has a density which is very close, and possibly exactly equal, to the critical density … we live in a flat universe. This is confirmed by recent observations of the size of fluctuations in the cosmic microwave background (CMB) radiation. (See below for details.)

Key concept

Astronomers currently believe that we live in a universe that has a density which is very closely equal to the critical density.

The *critical density* is the average density of matter in the universe which will allow the universe to expand forever, but such that that the rate of expansion will reduce to zero after infinite time. This is described as a 'flat universe'.

Describing and interpreting graphs showing the variation of the cosmic scale factor with time

- Figure 16.39 is similar to Figure 16.26 from the end of Section 16.3, but it has been annotated with references to critical density.
- A **flat universe** ($\rho = \rho_c$) is one in which the rate of expansion will reduce to zero after an infinite time.
- A **closed universe** ($\rho > \rho_c$) is one in which the density of the universe is higher than the critical density, so that 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** ($\rho < \rho_c$) is one in which the density of the universe is lower than the critical density, so that the universe will continue to expand forever.

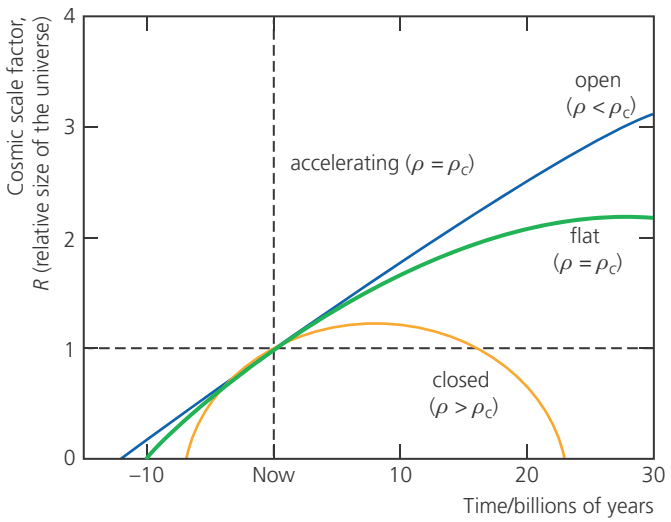


Figure 16.39

Key concept

Figure 16.39 is a cosmic scale factor, *R*–time sketch showing the possible effects *of different densities* on the evolution of the universe. The effect of *dark energy* has not been included (see below).

Describing qualitatively the cosmic scale factor in models with and without dark energy

- We have seen in Section 16.3 that evidence from type Ia supernovae indicates that the rate of expansion of the universe is expanding. This is shown in red in Figure 16.40.
- An **accelerating universe** cannot be explained by gravitation theory. An explanation in terms of *dark energy* has been proposed (see below).

Key concept

All recent evidence suggests that the rate of expansion of the universe is *accelerating*, despite having the necessary (critical) density for being gravitationally 'flat'.

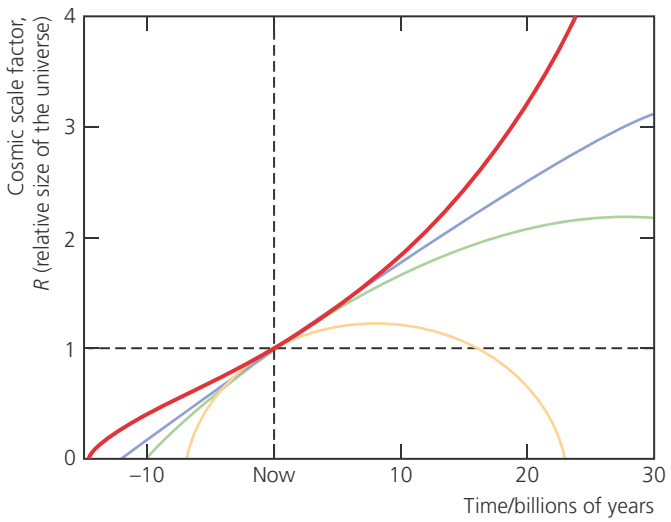


Figure 16.40

Common mistake

A gravitationally flat universe *can* also be accelerating. A flat universe is one in which $\rho = \rho_c$ and in which, *if there was no dark energy*, the rate of expansion would reduce to zero after infinite time. But there *is* dark energy and this is causing the rate of expansion to increase.

Dark energy

Revised

- Dark energy is considered to exist everywhere in space at very low concentrations. Its only interaction is gravitational and it is believed to account for about 68% of the mass-energy in the universe (see Figure 16.41).

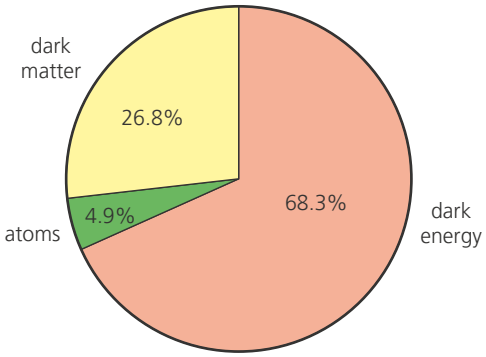


Figure 16.41

Key concept

Dark energy is an unknown form of energy which is considered to exert a negative pressure (like a repulsive force) on matter. The concept was developed to explain the expanding universe.

The universe cools as it expands

Revised

- As the universe expands, matter and energy spread out and the average temperature decreases.
- From Wien's law we know that $\lambda_{max} \propto \frac{1}{T}$ and we have already noted (Section 16.3) that the wavelength of radiation from soon after the Big Bang has become stretched to about 1 mm, equivalent to an average temperature of 2.76 K.
- As space expands $\lambda_{max} \propto R$, the cosmic scale factor, so that $T \propto \frac{1}{R}$. The average temperature of the universe multiplied by the cosmic scale factor is a constant.
- For example, when the observable universe was half its current size, $R = 0.5$ and $T = 5.52$ K.

Key concept

The average temperature of the universe, $T \propto \frac{1}{R}$.

Expert tip

Figure 16.42 represents the expansion of the universe. There are no details included that need to be remembered for examinations.

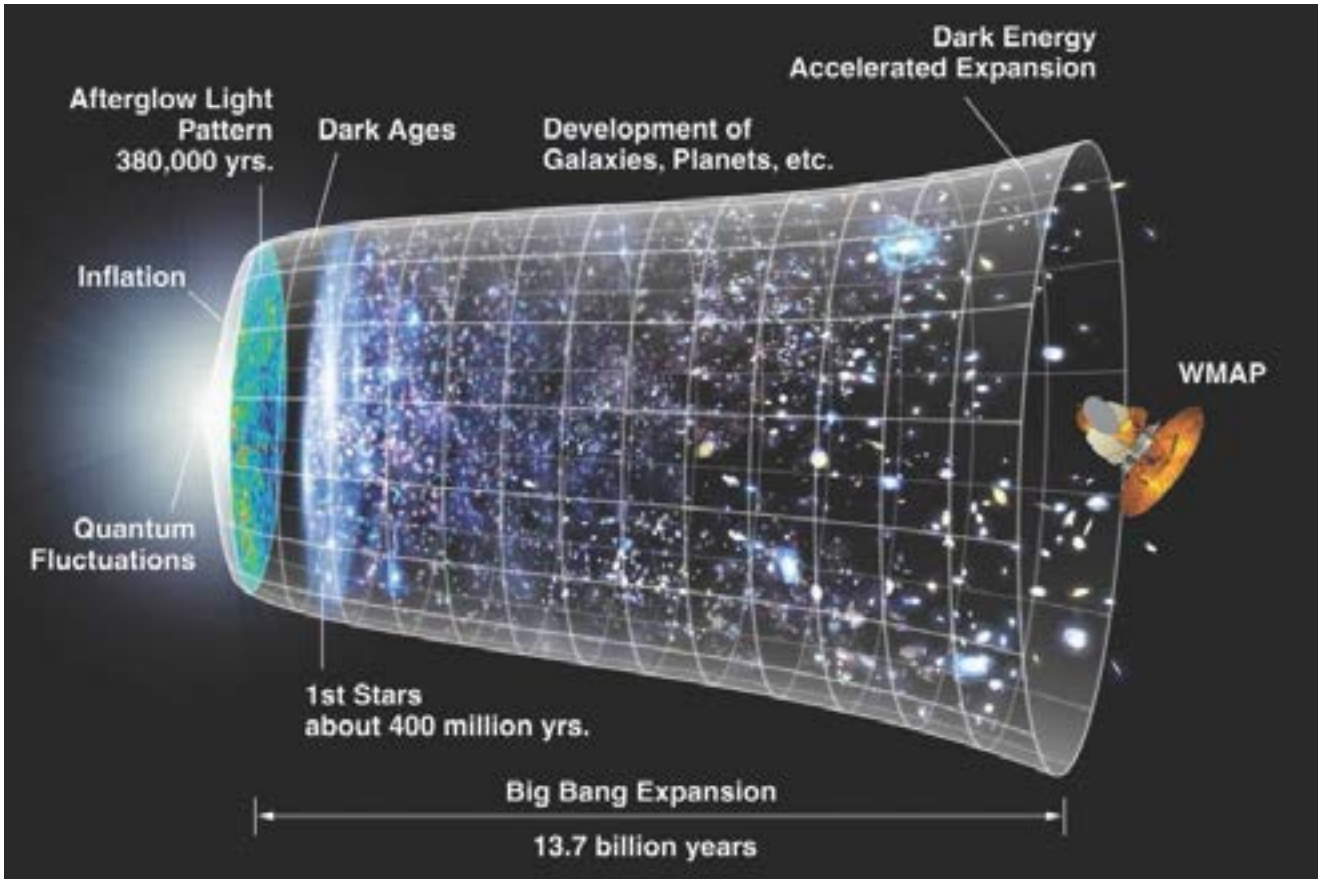


Figure 16.42

QUESTIONS TO CHECK UNDERSTANDING

86 a Explain the concepts of a flat universe and its critical density.

b Describe the fate of the universe (without dark energy) which is predicted by an average density which is

i just greater than

ii just less than, the critical density.

87 a Derive the equation $\rho_c = \frac{3H^2}{8\pi G}$.

b Confirm that using a value of $H = 73\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1}$ leads to a value of $\rho_c \approx 10^{-27}\,\text{kg}\,\text{m}^{-3}$ and a particle density of less than 10 atoms per cubic metre.

c What assumption did you make in answering the second part of (b)?

88 If the universe is expanding at an increasing rate, how will the cosmic scale factor change in the future?

89 Explain why astronomers needed to introduce the concept of dark energy after it was discovered that the universe was 'accelerating'.

90 Predict the average temperature of the universe:

a 14 billion years in the future

b 7 billion years in the future.

Fluctuations in the CMB

Revised

- In Section 16.3, it was explained that the same cosmic microwave background (CMB) radiation coming from all directions was evidence of an average temperature of 2.76 K and therefore evidence of the Big Bang.
- However, there are some *very small* fluctuations in the CMD, as shown in Figure 16.43.

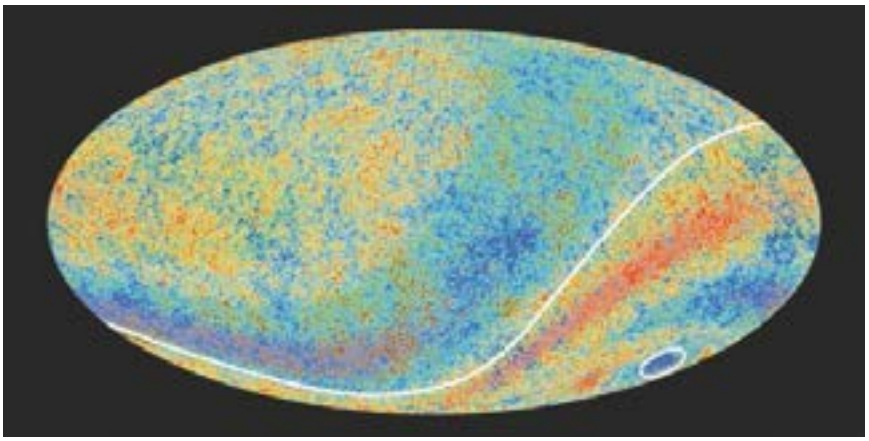


Figure 16.43

Describing and interpreting the observed anisotropies in the CMB

- The universe did not become transparent for radiation until it was about 400 000 years old. It was at this time that the CMB radiation was emitted.
- The *COBE*, *WMAP* and Planck Space Observatories have provided information about anisotropies, the latest estimates for the critical density and age of the universe, plus estimates of the proportions of observable mass, dark matter and dark energy in the universe.

Key concept

CMB is not 100% uniform from all directions. There are *very* small *fluctuations* (about 0.01% or less), called **anisotropies**, which have great significance.

Key concepts

The anisotropies in temperature provide important evidence about the early stages of the universe when the radiation was emitted, including information about tiny variations in density.

The **COBE**, **WMAP** and **Planck** Space Observatories have provided an ever-expanding bank of data on which astronomers are building an impressive understanding of the universe.

QUESTIONS TO CHECK UNDERSTANDING

91 The colours in Figure 16.43 have been added artificially. What do they represent?

92 If the latest measurements from WMAP predict a temperature of 2.725 K ± 0.04%, estimate the maximum absolute variation in detected temperatures.

93 Explain why fluctuations in CMB provide astronomers with information about the early development of the universe.

NATURE OF SCIENCE

Cognitive bias

Scientists must avoid confirmation bias, a tendency (often unintentional) to accept evidence in support of current beliefs and reject evidence to the contrary. The history of astronomy has many paradigm shifts (like an accelerating universe), all of which have received opposition from supporters of the accepted wisdom of that time.