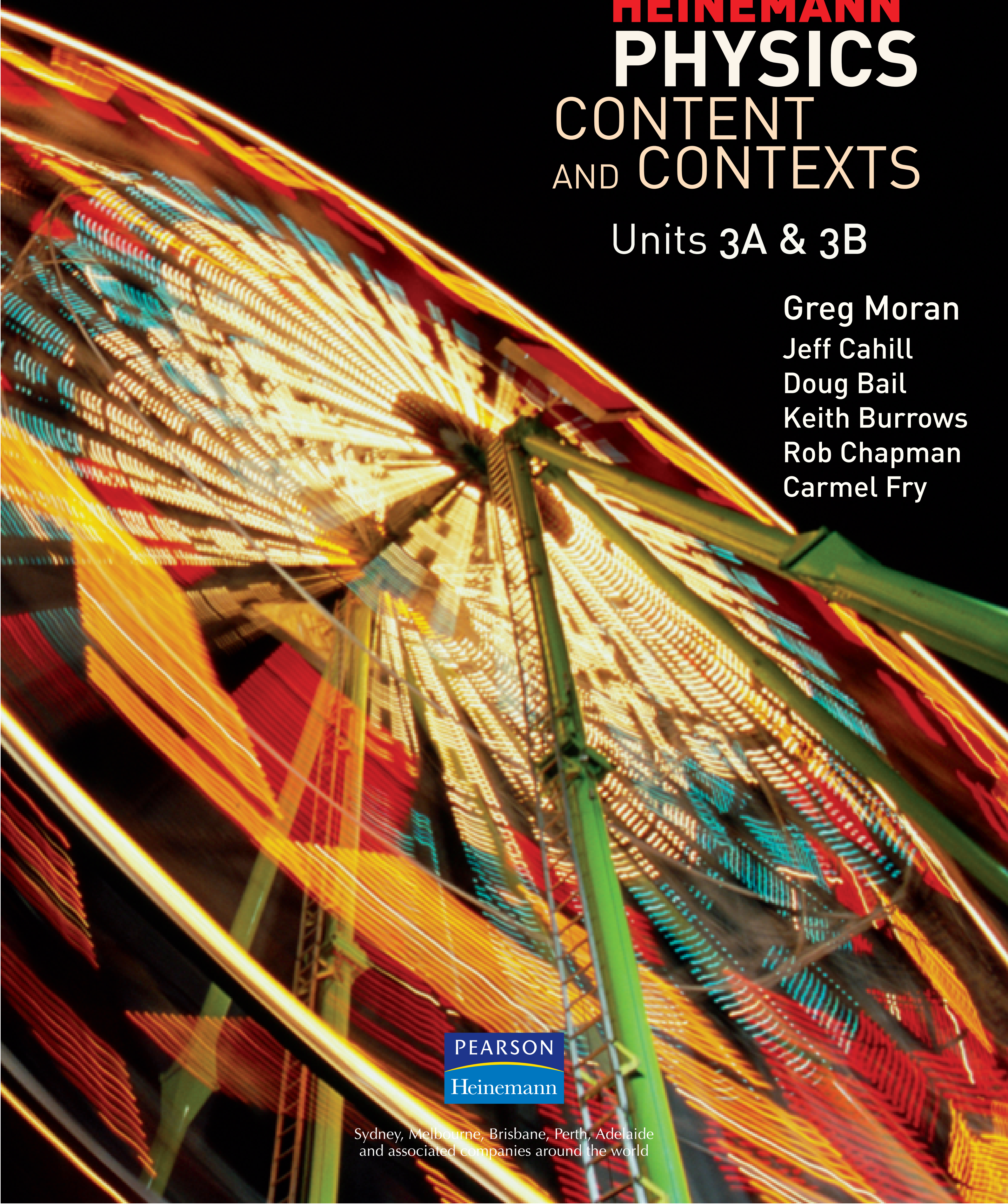
interactive CD included

HEINEMANN PHYSICS CONTENT AND CONTEXTS

Units 3A & 3B

Greg Moran
Jeff Cahill
Doug Bail
Keith Burrows
Rob Chapman
Carmel Fry



PEARSON
Heinemann

Sydney, Melbourne, Brisbane, Perth, Adelaide
and associated companies around the world

HEINEMANN PHYSICS CONTENT AND CONTEXTS

Units 3A & 3B

The complete package for the Western Australian Physics Course of Study

Heinemann Physics—Content and Contexts Units 3A & 3B is the most up-to-date and complete package for the Physics Course of Study in Western Australia. Physics is presented in meaningful contexts, and the unique structure offers teachers the flexibility to address the content in the way that suits the needs and interests of their students.

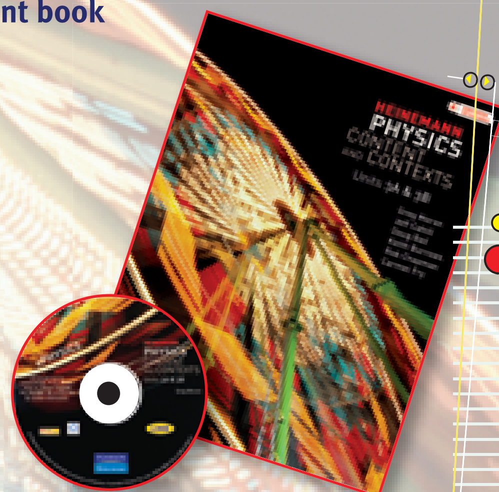
Heinemann Physics—Content and Contexts Units 2A & 2B is written to cover Stage 2 of the course of study and *Heinemann Physics—Content and Contexts Units 3A & 3B* covers Stage 3.

Heinemann Physics—Content and Contexts Units 3A & 3B student book

Includes ePhysics student CD

Key features:

- Specifically written for Stage 3 of the physics course in Western Australia
- Includes contexts and chapters
- Further contextual material embedded in the chapters
- Full colour design
- Structured to assist the understanding of key concepts
- Comprehensive range of questions at the end of each section and chapter

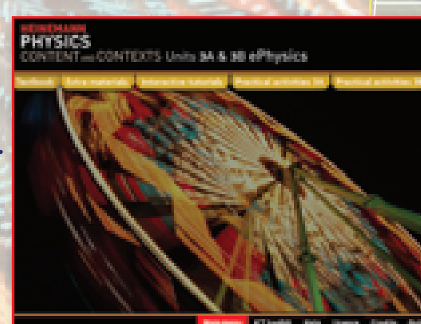


Each student book includes ePhysics, an interactive student CD containing:

- Electronic student book
- Extra contexts and content
- Interactive tutorials
- Practical activities
- ICT toolkit

Heinemann Physics—Content and Contexts Units 3A & 3B Companion Website

www.pearsonplaces.com.au



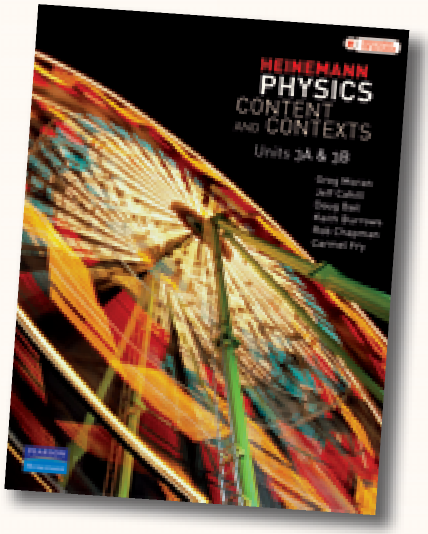
The Teacher Lounge at Pearson Places contains all the advice and support needed to readily and effectively implement the Western Australian Physics Course of Study. It includes:

- Teaching program and curriculum guide
- Worked solutions to all exercises in the contexts and questions in the chapters
- Practical safety notes for all activities on ePhysics

For more information on **Heinemann Physics—Content and Contexts Units 3A & 3B Companion Website** visit www.pearsonplaces.com.au

Introduction

Heinemann Physics—Content and Contexts Units 3A & 3B is part of the Heinemann Physics—Content and Contexts series, which has been written to meet the needs of Western Australian students and teachers, and has been created specifically to match the Western Australian Physics Course of Study to be implemented from 2010.



The authors have written a text that will support students' learning in physics while making the subject interesting, enjoyable and meaningful. The book uses clear and concise language throughout. All concepts have been fully explored, first in general and then developed in context. Illustrative material is fresh, varied and appealing to a wide range of students.

Each of the book's chapters has been divided into a number of self-contained sections. At the end of each section is a set of homework-style questions that are designed to reinforce the main points. More demanding questions are included at the end of the chapter. The large number of questions is designed to assess students' understanding of basic concepts as well as giving them practice at problem solving.

The chapter section questions are useful for both tutorial classes and homework assignments. A teacher might typically select some questions for immediate work in class and assign the rest as homework. There are over 600 questions in the text. Answers are supplied at the end of the text and fully worked solutions are available on the Teacher Lounge at Pearson Places.

Pearson Places www.pearsonplaces.com.au

The authors have retained many features of the previously highly successful textbooks for Year 11 and Year 12 physics. Within each section, the concept development and worked examples occupy the main column. The minor column has been set aside for some of the 500 photographs and diagrams, as well as small snippets of Physics file information. The longer pieces of high interest and context material are contained in the full-page width Physics in action sections. Both Physics in actions and Physics files are clearly distinguishable from remaining material, yet are well integrated into the general flow of information in the book. These features enhance students' understanding of concepts and context.

Newly featured in this series of texts are additional new areas that are entirely contextual. The contexts at the beginning of each unit cover a broad context that relate to the conceptual content of the course. The contexts include additional exercises and activities at appropriately challenging cognitive levels. Students and teachers choosing to cover a context first can be directed to the appropriate chapter sections when required. Alternatively, a more traditional approach including coverage of the content chapters prior to undertaking the context will enhance student learning due to the enrichment associated with additional exposure to applied and contextual material.

The textbook includes an interactive CD, ePhysics, which will enhance and extend the content of the texts. Included are:

- fully interactive tutorials that allow students to explore important concepts that may be too difficult, dangerous or expensive to do first hand in the classroom
- an innovative range of short and long practical investigations. These have been fully trialled and tested
- a complete electronic copy of the textbook
- an additional context for each of Units 3A and 3B

- a complete chapter of Working in Physics
- additional sections on Analysing motion and Understanding electromagnetism that revise and expand on important concepts from Stage 2 Physics
- an ICT toolkit with tutorials on spreadsheets, databases, Web use and more.

The *Heinemann Physics—Content and Contexts Units 3A & 3B* Teacher Lounge at Pearson Places supports the text and ePhysics and assists teachers implement, program and assess the course of study.

Acknowledgments

The publisher would like to acknowledge and thank the following writers who have contributed material to this text and interactive CD: Jeff Cahill, Doug Bail, Rob Chapman, Carmel Fry and Keith Burrows. The publisher wishes to thank Ross Fuhrmann for his contribution to this book. We also thank Gary Thomas for his thorough work in answer-checking this publication and for his reviews. George Przywolnik is also thanked for his reviews.

We would like to thank the following for permission to reproduce copyright material. The following abbreviations are used in this list: t = top, b = bottom, l = left, r = right, c = centre.

About.com Cars/Aaron Gold: p. 75 (inclinometer).

Alamy Limited: pp. 87br, 87tr, 449.

All Australian Nature & General Photo Library/ Northside Productions: p. 337t.

Australian Associated Press Pty Ltd: pp. 20, 25, 40, 45, 194, 244.

Australian Nuclear Science and Technology Organisation (ANSTO): p. 264.

Australian Synchrotron: pp. 350, 355.

Brookhaven National Laboratory, courtesy of Brookhaven National Laboratory: p. 339t.

Burrows, Keith: pp. 98, 337b, 454.

CERN: p. 345.

Connex: p. 113.

Coo-ee Picture Library: pp. 472br, 484l, 489.

Corbis Australia Pty Ltd: pp. 13, 34, 39, 165, 252b, 273, 281, 325, 362, 404t, 428, 437c, 437r, 471, 484c, 484r, 486.

Cross, Malcolm: pp. 96tr, 135, 139, 169, 408l.

CSIRO Publishing: p. 496; John Ward, p. 163.

David Malin Images: pp. 304b, 305, 307tl.

Dorling Kindersley: pp. 72r, 120.

ESO: p. 317 (b); Lars Lindberg Christensen, p. 317 (a).

European Space Agency (ESA): pp. 295, 322.

Flickr: Jeff Medaugh, p. 31; Abraham Orozco, p. 35.

Getty Images Australia Pty Ltd: pp. 183, 284, 307b, 328, 423b, 425, 439.

iStockphoto: pp. 19, 21, 28, 33t, 37, 48, 72l, 80, 84l, 84r, 86c, 86r, 148, 184b, 185, 202, 204, 242, 296, 298bl, 423tr, 437l, 438, 463, 512b, 513.

John Dolphin Photography: p. 83.

Lochman Transparencies: p. 164.

NASA: pp. 53, 61, 63, 70, 274, 302, 304t (all), 317 (c & d), 336tl; WMAP Science Team, p. 326.

Newspix: p. 167.

PASCO Scientific: p. 430tr.

Pearson Australia: Sarah James, p. 184t; Jacqui Liggett, p. 72c; Alice McBroom, p. 232tl.

Photolibrary Pty Ltd: pp. 64b, 66, 75b, 88t, 94, 121, 128, 129, 136, 142, 173, 231, 235, 245, 251, 256, 263, 267bl, 268, 269, 270, 271, 280, 287, 297, 298tl, 300, 301, 307cr, 311, 312, 316, 318, 334, 343, 404br, 429, 430bl, 474, 476, 478, 479.

The Picture Desk/The Art Archive/Archaeological Museum Bari/Gianni Dagli Orti: p. 156.

The Picture Source: pp. 99, 420.

Restrospect Photography/Dale Mann: p. 186.

Shutterstock: pp. 2–3, 4, 7, 15, 16, 17, 18, 86l, 87 (Gothic arch), 87 (Rheims Cathedral), 87tl, 87bl, 88b, 93, 96bl, 105, 119, 177, 192, 206, 210, 323, 401, 422, 439r, 495, 499, 512t.

Siemens: p. 149.

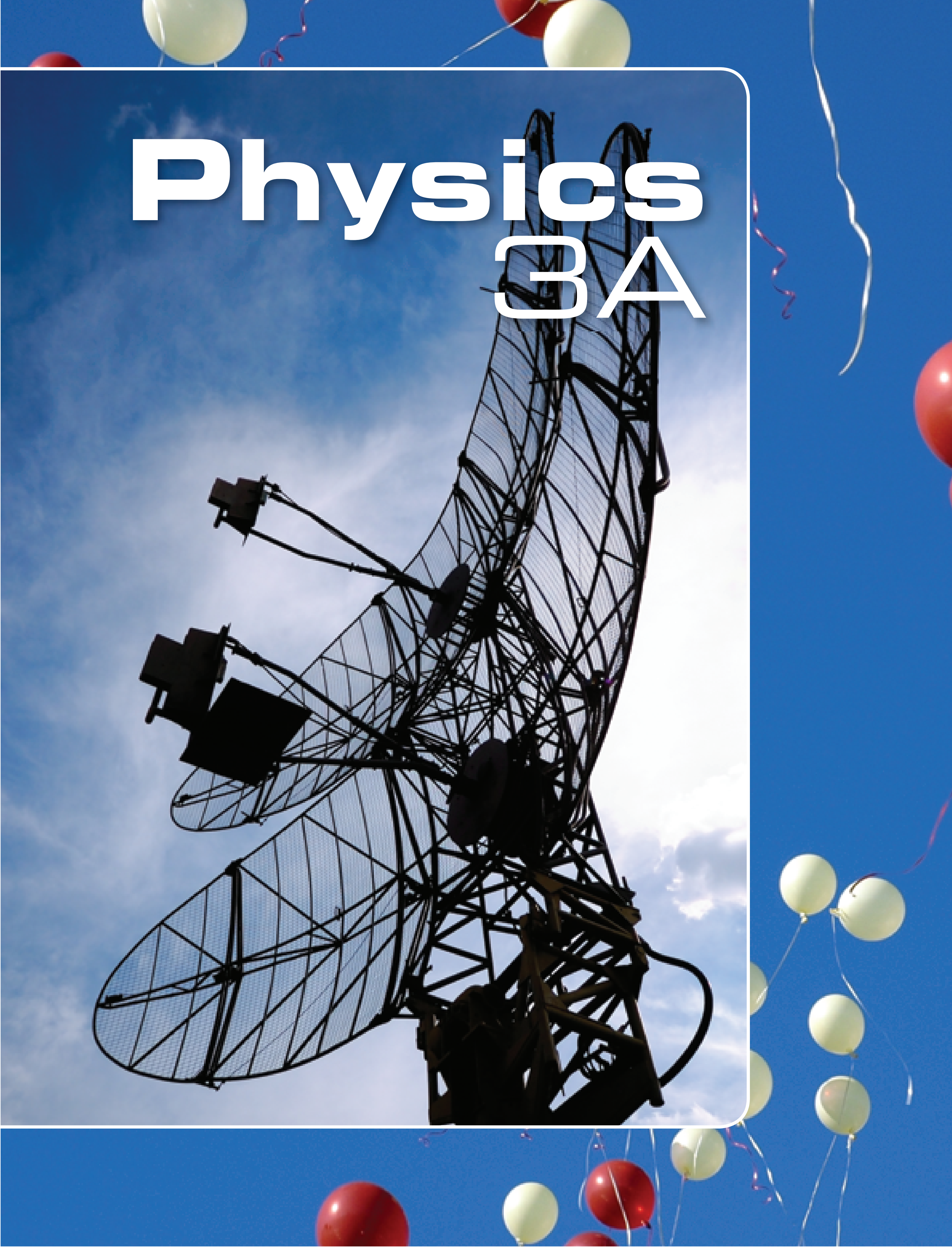
State Library of South Australia/Mountford–Sheard Collection: p. 307tr.

© STFC: p. 333.

www.gravity.uwa.edu.au, courtesy of www.gravity.uwa.edu.au: p. 267r.

Heinemann Physics Content and Contexts for Units 3A & 3B

Physics 3A



The first context section and content Chapters 1–4 cover the content required for the Physics course Unit 3A.

Outcomes

The Physics Outcomes are as follows:

Outcome 1: Investigating and communicating in physics

Students investigate physical phenomena and systems, collect and evaluate data, and communicate their findings.

In achieving this outcome, students:

- develop questions and ideas about the physical world to prepare an investigation plan
- conduct experiments and investigations
- analyse data and draw conclusions based on evidence
- evaluate the accuracy and precision of experimental data and the effectiveness of their experimental design
- communicate and apply physics skills and understandings in a range of contexts.

Outcome 2: Energy

Students apply understanding of energy to explain and predict physical phenomena.

In achieving this outcome, students:

- apply understanding of conceptual models and laws relating to energy
- apply understanding of mathematical models and laws relating to energy.

Outcome 3: Forces and fields

Students apply understanding of forces and fields to explain physical phenomena.

In achieving this outcome, students:

- apply understanding of conceptual models and laws relating to forces and fields
- apply understanding of mathematical models and laws relating to forces and fields
- apply understanding of the vector nature of some physical quantities.

It is envisaged that students will fulfil the requirements of Outcome 1 through investigative work in the classroom. Investigative work will typically be related to the content covered in Chapters 1–4. Working in Physics covers the basic requirements for understanding and processing measured quantities in the investigative work. A complete chapter that revises and expands on Working in Physics for Stage 3 can be found on the *ePhysics* CD.

Chapters 1–4 cover the required content for Outcomes 2 and 3.

Unit description

Unit 3A focuses on two broad areas of physics: ***Motion and Forces in a Gravitational Field*** (Chapters 1 and 2) and ***Electricity and Magnetism*** (Chapters 3 and 4).

In ***Motion and Forces in a Gravitational Field***, students learn about the motion of objects in one and two dimensions. They learn about the energy involved in motion, and the conservation of momentum. They study circular motion in horizontal planes as well as the effects of gravitational field forces on motion. They also learn about torque and how forces act on objects in equilibrium.

In ***Electricity and Magnetism***, Students learn about the concepts of charge and energy transfer in situations involving current electricity. They learn about magnetic fields and their interactions with current carrying conductors including the motor effect. They apply the concept of electromagnetism to electromagnetic induction and learn about the distribution of electrical power to consumers.

Contexts

The context material, preceding the content chapters, supports the content of Unit 3A:

- Fairground physics
- Power generation and distribution (CD).

Fairground physics

By the end of this context

you will have covered material including:

- applying the law of conservation of energy to the motion of a roller-coaster
- using the equations of motion to solve problems involving fairground rides
- exploring the concept of ‘g-force’ and apparent weight
- applying the concepts of circular motion in a horizontal plane to turns on a go-kart track and a carousel
- exploring the cost of running an electric motor to lift a roller-coaster up the first hill
- combining the effects of gravity and circular motion in a horizontal plane to a vortex ride
- modelling fairground rides to analyse circular motion
- applying the law of conservation of momentum to bumper car rides.

Wouldn’t it be great to go to a fun park like Adventure World or Disneyland for every physics class? Where every lesson was about something there: a ride, a machine, a drink dispenser. Well, without the application of physics principles, there would be very little fun in a fun park. Even sliding down a slippery slide at the local park involves balancing gravitational and frictional forces. During this context we will investigate the applications of physics in an amusement park or theme park, allowing you to analyse existing facilities or design new ones. You will gain a better understanding of the tremendous effort that has gone into the design of even fairly simple rides.

Figure fp.1 Fairground rides depend on physics for their safe construction.

Carousel

A carousel (or merry-go-round) is a simple application of circular motion and centripetal force. The simplest form has some horses mounted onto a large rotating platform. More complicated models have the horses moving up and down as they go around.

CENTRIPETAL FORCE is the force that acts on a body at 90° to the motion of the body, such that it causes the body to move in a circular path. The direction of the centripetal force is towards the centre of the circular path.

Since this ride is typically for the younger child, the main consideration is to make sure that the carousel does not spin so fast as to throw the children off. (Note again that 'throw' is not the right word.) The children have mass and are moving (on the carousel), and Newton's first law says that they will maintain constant velocity (speed and direction) unless acted on by a force. If the child lets go, they will continue at their velocity—in a straight tangential line—and not follow the curve of the carousel.

NEWTON'S FIRST LAW states that an object will continue with its velocity unless acted upon by an external, unbalanced force. So if an object is stationary, it will remain stationary; if it is moving, it will continue to move unless a force causes it to accelerate.

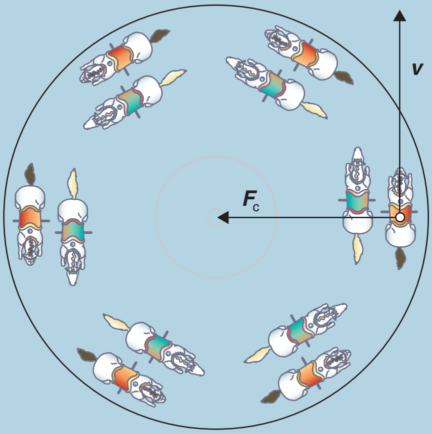


Figure fp.2 Maintaining constant velocity (speed and direction) if flung from a carousel

In order to turn in the circle, a force must act to change the direction of the riders' motion; it must push on them in a direction that is towards the centre. On a carousel, this involves a sequence of forces that is involved with the child holding onto the horse, the horse being fastened to the turning platform, and the platform being strong enough to take the strain.

If we assume that the weakest link in that sequence is the child holding on, we can calculate a maximum safe speed for the carousel. There are two approaches. We could estimate the maximum force that a small child can comfortably use to hold on for a long period of time, and use this to calculate the circular velocity given the radius of the ride. Alternatively, we could estimate the maximum desirable impact force should a child fall off, and use this to calculate a maximum velocity. Ideally, we should use both approaches and take the lowest value as the safe speed.

See 1.2 Circular motion in a horizontal plane on page 33.

Since a carousel turns in a circle, the ***CENTRIPETAL FORCE*** equation applies:

$$\boldsymbol{F}_c = \frac{mv^2}{r}$$

where $\boldsymbol{F}_c$ is the centripetal force (N), m is the mass of the object undergoing the circular motion (kg), v is the speed at which the object is moving (m s^{-1}) and r is the radius of the circular path (m).
When calculating the maximum safe velocity, this equation rearranges to give:

$$v_{max} = \sqrt{\frac{\boldsymbol{F}_c r}{m}}$$

The mass, m, in this equation is the mass of the child riding the carousel.

Note that the value of $\frac{\boldsymbol{F}_c}{m}$ in this equation is also the acceleration, so the equation becomes:

$$v_{max} = \sqrt{\boldsymbol{a}_{max} r}$$

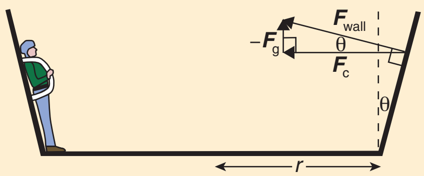


Figure fp.3 There are a variety of rides that are based on being spun very fast.

The use of a maximum acceleration allows a general answer, instead of having to estimate a series of forces and masses for the children. Given that a car accelerates at less than 0.100***g*** for normal accelerations, a reasonable maximum acceleration for our carousel horse would be less than this. In a car a child has a seat belt, but there are no seat belts on a carousel!

When calculating the force to be withstood by the attachments holding a horse onto a carousel, or restraints on any other ride, safety standards usually require that the actual force that can be withstood must be several times larger than what is predicted. This is called a safety factor. A safety factor of 10 means 10 times the expected force must be able to be withstood.

Exercises

E1 The designer of a small child's carousel with a diameter of 6.00 m would like the maximum acceleration experienced by the child to be less than 0.0200***g***. What is the maximum speed it should turn at?

E2 A very young relative wants a ride on a carousel with three columns of horses. If the main consideration is that they might not be able to hold on very well, should you recommend that the relative ride on the:

A inside column?
B middle column?
C outside column?

E3 On the same carousel as in exercise E2, an older and stronger relative wants a fast ride. What would you recommend for them?

E4 In designing the attachment of a 20.0 kg horse onto a carousel designed to carry riders up to 50.0 kg mass, and at a centripetal acceleration of 0.0500***g***, what force does the attachment have to be able to withstand, if local regulations require a safety factor of 12?

Spin dry cycle!

For older children, there might be more excitement in being spun even faster than a carousel. There are a variety of rides based on the principle of spinning very fast. Using the same equations as for a carousel, we could design a drum, in which children could stand, and be spun, like being inside a spin drier that spins in a horizontal plane. They would feel as though they were being pushed against the wall of the drum. But we know that the wall is actually pushing against them, causing their straight line inertia to be changed to fit the circular path of the drum. If the spin was fast enough, the centripetal force added to the gravitational force could make it seem as though the wall of the drum were the floor.

Alternatively, chairs hung from chains could be spun. The resultant inertial effect would have the chairs being lifted off from vertical at an angle determined by the speed of spin and the radius of curvature.

In both cases, the force exerted by the chair or wall on the rider is the vector sum of the centripetal force and the opposite of the gravitational force.

The calculation of the force experienced by the rider depends on the mass of the rider, so again it is easier to calculate the acceleration in g's, which is independent of mass. The acceleration experienced by the rider is the centripetal acceleration:

$$a = \frac{v^2}{r}$$

The angle for the acceleration is the same as that for the centripetal force; that is, towards the centre.

Physics file

A spin dryer uses Newton's first law to dry clothes. There is a rotating drum in a spin dryer with many small holes drilled into it. When it is filled with wet clothes and begins to spin the walls of the drum provides the force that changes the motion of the clothes. The water is more mobile than the clothes and is able to continue its straight line motion by moving through the holes. Perhaps it would be better to say that the holes move through the straight line of motion of the water.

Physics file

This type of arrangement can be used in space stations to provide a gravity-like environment. A spinning circular section of the space station could allow astronauts to walk with apparent weight on the inside walls of the circular section. In a famous scene from the movie *2001: A Space Odyssey*, an astronaut is seen jogging in a circular path on the walls of the space ship. The apparent weight is due to the force of the walls on the feet of the jogger, who would continue in a straight line if the walls weren't there.

Exercises

E5 What centripetal force due to the wall will be experienced by a 60.0 kg rider in a drum of radius 5.00 m spinning at 4.00 m s^{-1}?

E6 If a spinning drum ride cannot exceed a radius of 10.0 m, and the riders are not to experience a centripetal acceleration due to the wall on the rider of greater than 2.00$\boldsymbol{g}$, what is the maximum speed at which the drum can spin?

Roller-coasters

A roller-coaster car depends on the action of gravity for movement over most of its journey. The car is hauled to the highest point of the track, and then released, to roll down to the bottom. Along the way, bumps, loops and turns make the ride more exciting. The length of the ride will depend on two factors: the height to which the car is hauled at the start, and the friction between the car and the track.

Figure fp.4 A roller-coaster relies on the force of gravity for its exhilaration.

See 2.1 Gravitational fields on page 54.

Rolling down a hill

Let's start by looking at a very simple ride. The car is hauled to the top of a hill, and rolls down a straight track to the bottom. At the top of the hill, the stationary car has gravitational potential energy.

The ***GRAVITATIONAL POTENTIAL ENERGY EQUATION*** states that the work done to raise an object in a gravitational field is:

$$E_p = m\boldsymbol{g}\boldsymbol{h}_y$$

where m is the mass of the car (kg), $\boldsymbol{g}$ is the gravitational acceleration (9.80 m s^{-2}) and $\boldsymbol{h}_y$ is the vertical height (m).

The vertical height is with respect to some reference point, typically the surface of the ground underneath the car. Any reference point can be chosen as long as all calculations within a situation relate to the same reference point. It simplifies calculations to choose a reference point so that the gravitational potential energy at some part of the situation is zero and the rest is positive; that is, choose a reference point at the lowest part of the track. In this case, choosing a reference point at the lowest point of the track allows us to ignore the gravitational potential energy at that point, and therefore the vertical height in the equation is the height above the lowest point of the track.

As the car starts to move down the slope, it will accelerate due to the force of gravity. The vertical or y acceleration will be equal to $-\boldsymbol{g}$, but since the car is not falling in the y direction, it will gain horizontal or x velocity as well as y velocity. These velocities can be determined by the equations of motion, once we know the elapsed time (Δt) and the angle of the slope away from the vertical θ.

Physics file

In this text, vertical and horizontal vectors will be described using the x,y,z-coordinate convention, where x is horizontal and in the plane of the page, y is vertical and in the plane of the page, and z is into and out of the page.

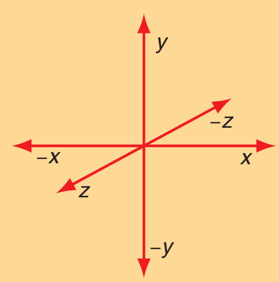


Figure fp.5 The x,y,z-coordinate convention

See 1.4 Vectors and free-body diagrams on page 416 on CD.

See 1.5 Motion in a straight line on page 422 on CD.

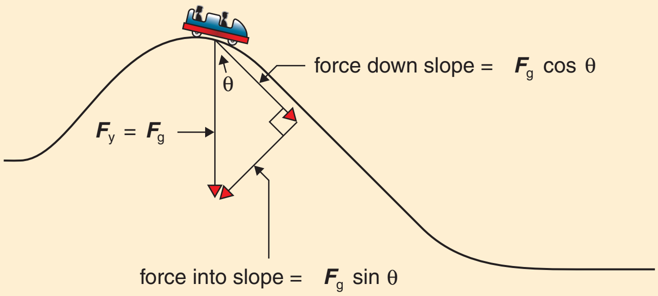


Figure fp.6 The down-slope and into-slope components of the force acting on a roller-coaster car travelling down the slope

Down the slope, the force and acceleration can be calculated by:

$F_{g\ \text{down slope}} = F_g \cos\theta = mg \cos\theta$

$a_{g\ \text{down slope}} = g \cos\theta$

Into the slope, the force can be calculated by:

$F_{g\ \text{into slope}} = F_g \sin\theta = mg \sin\theta$

NEWTON'S THIRD LAW states that for every action (force) there is an equal and opposite reaction (force).

APPARENT WEIGHT is the weight force that you feel. In a stationary lift you feel as if you are pressing on the floor with your normal weight force. As the lift accelerates upwards, you feel the sensation that you are pressing on the ground with more force; as the lift accelerates downwards, you get the feeling that you are not pressing on the ground as hard. Your apparent weight can change depending on the acceleration of the surface on which you are standing.

The force into the slope is countered by the track; otherwise, the car would sink into or fall through the track. According to Newton's third law, the force of the car on the track is balanced by the equal and opposite reaction force of the track pushing back on the car. The force of the car into the surface of the track provides the apparent weight of the car. The reaction force of the track on the car balances the into-the-slope component of the force due to gravity on the car and therefore they cancel each other out, resulting in no acceleration of the car into the slope.

Physics file

The equations of motion can be broken into two groups, those involving constant motion and those involving acceleration. Constant motion equations:

$$v_{av} = \frac{d}{\Delta t}$$

Accelerating motion equations:

$$v_{av} = \frac{v + u}{2}$$

$$a = \frac{v - u}{\Delta t}$$

$$s = ut + \tfrac{1}{2}a\Delta t^2$$

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

See Vector components on page 418 on CD.

The **LAW OF CONSERVATION OF ENERGY** states that energy can neither be created nor destroyed; it can only change form.

Given a time interval, and the angle of the slope, and the length of the slope, it is possible to calculate the fastest speed that the car will reach, but there is an easier way. By using the principle of conservation of energy (which states that energy cannot be created or destroyed), we can say that:

$$E_{p,\ \text{at the top}} = E_{k,\ \text{at the bottom}} + E_{\text{lost through friction, sound, heat or wear}}$$

See Energy and momentum on page 433 on CD.

KINETIC ENERGY is the energy an object has due to its motion.

Interactive tutorial

Kinetic and gravitational potential energy

Calculating the energy lost through friction from a theoretical perspective is very difficult, but it will be shown later that it is very easy to measure (see the following section '. . . and up the other side'). We can calculate the kinetic energy, $E_k = \frac{1}{2}mv^2$, so, for the time being, ignoring the energy losses from friction, we can say that:

$$E_{p\,lost} = E_{k\,gain}$$

$$mg\Delta h_y = \tfrac{1}{2}m\Delta v^2$$

Since the masses cancel,

$$g\Delta h_y = \tfrac{1}{2}\Delta v^2$$

$$\Delta v^2 = 2g\Delta h_y$$

$$\Delta v = \sqrt{2g\Delta h_y}$$

This equation is just a special case of one of the equations of motion:

$$v^2 = u^2 + 2a\Delta s$$

where $u = 0$ m s^{-1}, $a = g$ m s^{-2} and $\Delta s = \Delta h_y$ m.

*

See Equations of motion on page 431 on CD.
See Energy on page 433 on CD.

This velocity represents the greatest velocity possible for an object that drops a vertical height, Δh_y. Interestingly, the motion down the slope is independent of mass. The gravitational force acting on a larger mass is greater, but the greater mass requires a larger force to achieve the same acceleration, so these two effects cancel each other out. The exact path (the slope or shape of the track) does not seem to matter either. Where the shape and slope of the pathway will be important is in considering how much friction is encountered. Friction will slow down the object, so more friction because of a longer path should mean a slower car at the end of the ride.

Practical activity

8 Conservation of energy

Physics in action — A persistent misconception

One of the most persistent misconceptions in physics is the idea that gravity will cause a larger mass to drop at a faster rate than a smaller mass. This idea has its origins in our experience of feathers floating gently down, as opposed to a rock dropping to the ground. We can feel the difference in the mass of these two objects and see them dropping at different rates. However, we do not consider the effect of air resistance on the objects. If we could experience these two objects falling without the effect of air resistance we would see them falling at the same rate. But this does not mean that they both experience the same force due to gravity. In fact, in order to fall at the same rate, a larger mass experiences a greater force than the smaller mass experiences. From Newton's second law, the acceleration is equal to the ratio of the force to the mass, $a = \frac{F}{m}$. If the ratio for the two objects is the same then they will both have the same acceleration due to gravity. So an object with 10 times the mass will experience 10 times the force; the ratio will, however, be the same. The best way to think about the force due to gravity is to consider that the force acts on each particle of matter in each object; the total force is the sum of all the little forces on each particle of matter. If there are more particles in a more massive object, then there will be more total force acting on it. If there are 10 times the number of particles then there are 10 times the total force acting on the mass.

✓ Worked Example fp.1

What is the theoretical maximum speed of a roller-coaster car starting at a height of 75.0 m vertically above the lowest point of its track?

Solution

$\mathbf{g} = 9.80 \text{ m s}^{-2}$

$\Delta \mathbf{h}_y = 75.0 \text{ m}$

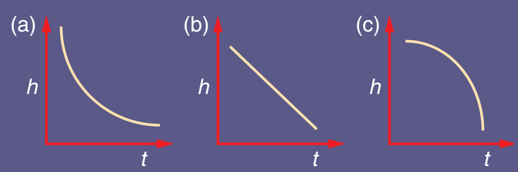
$$v = \sqrt{2\mathbf{g}\Delta\mathbf{h}_y}$$
$$= \sqrt{2(9.80)(75.0)}$$
$$= 3.83 \times 10^1 \text{ m s}^{-1}$$

Hence, the theoretical maximum speed is 38.3 m s^{-1}. Note that there is no need to include directions with the vectors as the answer is speed, which is a scalar quantity.

See Acceleration on page 426 on CD.

Investigation

There are three general types of paths possible for the roller-coaster: (a) exponential (dropping steeply and then levelling off), (b) a straight line, and (c) another exponential (dropping slowly, and then steeply).



These have different acceleration profiles and result in different time intervals to get to the same point. Obviously, the straight line is the shortest distance to travel, but it is not the quickest way down! See if you can figure out which will be the quickest, and why. Confirm your answer with toy cars and plastic track. Time how long it takes a toy car to reach a certain point, not how long it takes to travel the length of a section of track.

The equations show that the greatest possible velocity depends on the height from which the object starts dropping or rolling. Therefore, we have found that the higher the start of the ride, the faster it will be going at the bottom. Sometimes it is good to realise that physics can prove something that is so obviously true!

Theoretically then, our well-oiled roller-coaster track can have as many loops, bumps and turns as we like without affecting how fast the roller-coaster car will be going at the bottom. Realistically, though, the shape and length of the track will affect the duration and size of the frictional force, which will slow the car down. Therefore, loops, bumps and turns will all act to slow the car down, and make for a less interesting speed thrill. A straight drop down will give a better speed thrill, but has less interest. A reasonable compromise would be to drop the car from the top to the lowest point straight away, to gain the maximum speed thrill, and then go into the turns and loops. This large drop before the turns and bumps is a common design feature of roller-coasters. Look for it next time you ride one.

. . .and up the other side

Let's make the track a little more complicated: the shape of a U. How far will the car go up the other side? The kinetic energy at the bottom of the U is equal to the potential energy that the car had at the start (at the top of the U). Since energy cannot be created or destroyed, the kinetic energy gets converted back into gravitational potential energy as the car moves up the other side and slows down. The height that the car reaches on the other side should equal the height from which it started, but it won't. The reason for this is that some energy is lost to the system through friction: energy is being given out as heat or sound, or wear. Think about the rumbling sound a roller-coaster makes; a large amount of energy is being wasted to produce that sound.

This gives us a way to measure the energy lost through friction. Since:

$$E_{\text{p, at the top}} = E_{\text{k, at the bottom}} + E_{\text{lost through friction, sound, heat or wear}}$$

$$E_{\text{p, at the top}} = E_{\text{p, at the high point of other side}} + E_{\text{lost through friction, sound, heat or wear}}$$

then

$$E_{\text{lost through friction, sound, heat or wear}} = E_{\text{p, at the top}} - E_{\text{p, at the high point of the other side}}$$

so

$$E_{\text{lost through friction, sound, heat or wear}} = m\boldsymbol{g}\Delta\boldsymbol{h}_y$$

where $\Delta\boldsymbol{h}_y$ is the difference in vertical height between the start of the ride and the high point on the other side

This gives us a way of measuring the energy lost through friction and might be of interest, but it is not of practical use in designing a roller-coaster, since it can only be used after the coaster is built. Understand, though, that a roller-coaster ride can have hills and bumps, but each high point encountered on the ride must be lower than the previous high point to compensate for the energy loss due to friction. The calculations of theoretical frictional energy loss are too complex to investigate here. In your own design of a model roller-coaster, make what you think is a reasonable adjustment for friction by lowering the maximum heights by a certain percentage each time. If you are wrong in your guess, it is not too difficult to adjust a model (at least compared with a real coaster!)

See Conservation of energy on page 435 on CD.

✓ Worked Example fp.2

A 500.0 kg roller-coaster at a fairground starts 50.0 m above the ground, goes down into a dip, and just manages to roll over the next hill that is 42.0 m above the ground. How much energy has been lost through friction?

Solution

$m = 500.0$ kg

$\boldsymbol{g} = 9.80$ m s^{-2}

$\Delta\boldsymbol{h}_y = (50.0 - 42.0)$ m

$$E_{\text{friction}} = E_{\text{p, start}} - E_{\text{p, end}} = m\boldsymbol{g}\Delta\boldsymbol{h}_y$$

$$= (500.0)(9.80)(8.0)$$

$$= 3.92 \times 10^4 \text{ J}$$

Hence, the energy lost through friction is 39 200 J or 39.2 kJ.

Investigation

A roller-coaster rolling backwards and forwards through a U-shaped section of track appears to have very similar motion to that of a pendulum. Design and conduct an experiment to determine whether or not a pendulum and a roller-coaster car move in the same way.

Exercises

E7 If a roller-coaster car has a mass of 1.00 tonne including the passengers, what force will be accelerating it down a track that is 50.0° down from horizontal?

E8 A roller-coaster car is rolling down a slope that is 60.0° up from vertical. If the component of gravitational force acting on the car down the slope is 500.0 N, what is the mass of the car?

E9 What is the theoretical maximum speed of a roller-coaster car starting at a height of 50.0 m above the lowest point of its track?

E10 A 900.0 kg roller-coaster car starts its run 75.0 m above the ground, and rolls to a height of 62.0 m up a smooth U-shaped track before slowing to a stop. How much energy has been lost due to friction and other causes?

E11 The owner of a fairground wants a roller-coaster that will travel at 100.0 km h^{-1} at the bottom of the first hill. How high will it need to start?

E12 Will you travel faster, slower or at the same maximum speed if you are the only person on a roller-coaster ride, compared with sharing the ride with nine other people? (Ignore the effect of friction.)

E13 A roller-coaster car is approaching a rising slope at 19.0 m s^{-1}. What vertical height can it travel up the slope, assuming no energy is lost through friction?

E14 Assume that each energy transformation (e.g. kinetic → potential) loses 10% to friction. (This is an oversimplified, non-realistic assumption.) On a sheet of graph paper, accurately represent a sideways view of a roller-coaster track designed so that each successive hill is at the maximum height to allow the coaster car to just roll over it.

Acceleration is fun?

The excitement of a roller-coaster is also due to the unusual accelerations experienced during the ride, producing apparent forces from weightlessness to increased weight, and being thrown sideways. These accelerations have effects on the human body, depending on the direction in which they act.

A person's blood is only loosely held in the body; think of the body as a bag of water. When the body is accelerated upwards, the blood will tend to stay where it is (as defined by Newton's first law on inertia). Vertical acceleration upwards can therefore cause the blood in a person's body to drain away from their brain and into their feet. This can cause a blackout (a temporary loss of consciousness).

Vertical acceleration downwards (not just falling, because the blood will accelerate down at the same time as the body during a fall, but, for example, during a power dive in a plane) causes the opposite effect: the brain becomes oversupplied with blood, and a 'red-out' can also result in loss of consciousness. A person may also vomit when experiencing large downward accelerations because the stomach contents will also tend to stay in place due to inertia, as their body accelerates downwards.

If vertical acceleration in either direction is continued for a long time, death can result because a 'red-out' can cause bursting of blood vessels in the brain and a blackout can allow brain cells to die due to lack of oxygen. Some people are more susceptible to these effects of acceleration than others. The NASA astronaut training program tests people in special machines to see if they can tolerate these extremes of acceleration before allowing them to continue training.

The human body is better able to cope with horizontal acceleration, where the blood flow up and down the body is not affected. Up to 4g (where g is the gravitational acceleration, 9.80 m s^{-2}) can be tolerated by most people, and up to 10g can be tolerated for very brief periods. For comparison purposes, a car accelerating from 0 to 100 km h^{-1} in 6 s is accelerating at 0.47g. To accelerate at 4g, a car would go from 0 to 100 km h^{-1} in just 0.7 s!

Physics file

Weightlessness is a situation in which either you are not in contact with a surface and are falling due to gravity, or the surface that you are nearest to is falling at the same rate as you are due to gravity. When you are standing on a surface and it is not moving relative to you, then the force due to gravity that you apply to the surface is your weight force. Weight is measured in newtons (N). You can feel weightless even though there is gravity acting on you. There are places in deep space where the force due to gravity is close to zero; you would feel weightless there too, but you would not be falling.

*

See Acceleration on page 426 on CD.
See Gravitational fields on page 58.

In either horizontal or vertical acceleration, people who are not healthy are more likely to suffer ill-effects. There is also a risk with sudden accelerations of falling against an object, or even suffering whiplash, where an injury occurs to the neck because the neck muscles cannot hold the head upright against the forces involved. If you have ever stood in a swerving bus, you would be aware that even very low accelerations (typically less than 0.1***g*** in a bus) could cause you to lose balance and stumble or fall. Imagine what it would be like if the acceleration were 100 times greater!

Force is a vector, so the resultant force felt by the person on the roller-coaster is the vector sum of all the forces on the person. Of course, one of these is always the gravitational field strength, which operates down, and is equal to 9.80 N kg^{-1}. To this we can add whatever other forces are being experienced by the person.

NEWTON'S SECOND LAW states that the accelerating force on a body is equal to the rate of change of momentum of the body. The equation describing this can be manipulated to become force = mass × acceleration.

The person feels their weight because the gravitational force pulls them down onto a surface, but the surface material, on which they are standing, resists any motion in that direction. It effectively pushes back up. That upward push is usually equal and opposite to gravity, so the two cancel each other out, and there is no resultant vertical acceleration or motion. It is this upward push that gives the sensation of weight, even though the gravitational force is downwards.

The ***ACTION FORCE*** is the initiating force, the force that is delivered by the actions of a person. The ***REACTION FORCE*** occurs equally and oppositely due to the action force. If the action force did not exist, then there would be no reaction force.

In free-fall, the situation is a little more complicated. The person feels weightless. (Note that there is still gravity; it is never correct to say that there is no gravity.) You have probably felt a similar 'lessening of weight' when standing in an elevator as it starts to go down. There is still the force of gravity when free-falling, but since there is no surface to oppose it, there is no resistance, and therefore no *sensation* of weight.

This feeling of weightlessness can cause severe nausea in some people, so free-falling is not recommended for all rides. On the roller-coaster, it is not likely that the design will permit a true free-fall situation because of the action of friction. In any situation where the car is rolling down a slope, there will be some friction between the car and the track, no matter how steep it is (as long as it is not vertical). There will also be the friction due to air resistance. These opposing forces combine to reduce the force down the slope due to gravity, because friction always acts in the opposite direction to the motion.

Physics file

Pilots use flight suits with in-built compression chambers that inflate and squeeze the legs and lower body during a manoeuvre that could cause a blackout. This forces the blood out of the legs and allows more blood to be available for the pilot's brain.

Figure fp.8 Astronauts are tested in a g-force simulator before being allowed to go on space missions. The astronaut is seated in the chamber at the end of the spinning arm.

See Momentum on page 438 on CD.

Figure fp.9 When accelerating vertically, we can gain a sense of reduced acceleration because the force of air resistance acts against the force of gravity, and the resultant downwards force is less than the force of gravity alone.

Practical activities

9 Newton's second law
10 Newtown's second law II
11 Action and reaction

Interactive tutorial

Braking (video analysis of braking)

Therefore, as the car starts to drop, the riders will feel less heavy as they accelerate downwards. This explains the stomach-in-the-mouth feeling as the roller-coaster begins to drop, typically resulting in screams of terror. (We learn at a young age that falling is painful, so the sensation of falling produces a fear response.)

When the car reaches the bottom of a dip in the track, and starts moving up the other side, it is accelerating down the slope while moving up the slope. The down-slope acceleration is due to the force of gravity, and will slow the car down as it goes up the slope on the other side. However, since the direction of movement in the y direction (vertical velocity) of the car has changed from downwards to upwards, there must have been an upwards acceleration of the car at some stage in the ride, and therefore there must be a force acting upwards. This force is supplied by the track. Since the car is therefore accelerating upwards due to the direction change, the people in the car must also accelerate upwards against their inertia. In the roller-coaster car, the riders experience this as an increase in the reaction force produced by the car pushing them upwards as this accelerating force is added to the force of gravity, and so the riders feel heavier. Similarly, when the direction of movement in the y direction (vertical velocity) of the car changes from upwards to downwards, there must have been a downwards acceleration of the car during this stage of the ride. In the roller-coaster car, the riders experience this as a reduction in the reaction force produced by the car on them, and so the riders feel lighter.

An alternating series of bumps and dips (hills and valleys) will cause the riders to experience alternate sensations of lightness and heaviness. The degree of sensation will depend on the vertical acceleration, which will depend on the steepness of the track. The change in direction from upwards to downwards, or vice versa, divided by the time it takes to do so, is equal to the acceleration experienced. A shorter time interval will result in greater accelerations. Having steeper hills or valleys in the track can create a shorter time interval, and greater accelerations. Therefore, a compact track design will result in more intense sensations than a long gently sloping one.

✓ Worked Example fp.3

A roller-coaster car is approaching the top of a hill in the track, 15.0 m lower than the summit. It is travelling in the y direction at 10.0 m s^{-1} (vertically upwards). Another 4 s later, it is travelling in the y direction at –12 m s^{-1} (vertically down) and is 15.0 m down the other side of the hill. What average acceleration do the riders experience during this time interval?

Solution

↕ +/−

$\boldsymbol{u}_y = +10.0 \text{ m s}^{-1}$

$\boldsymbol{v}_y = -12.0 \text{ m s}^{-1}$

$\Delta t = 4.00 \text{ s}$

$$\boldsymbol{a}_y = \frac{\boldsymbol{v}_y - \boldsymbol{u}_y}{\Delta t} = \frac{(-12.0) - 10.0}{4.00} = -5.50 \text{ m s}^{-2}$$

Hence, the riders experience an average acceleration of 5.50 m s^{-2} downwards.

Safety

Even though collisions are frowned upon by the operators, crashes do happen on the go-kart track. Inertia means that, while a go-kart is braking, the bodies of any passengers will want to continue to move at a constant speed in a straight line. This can have unfortunate results for the passengers if the deceleration is sharp and sudden.

See Momentum on page 439 on CD.
See Impulse on page 440 on CD.

INERTIA is a property of a body that is related to mass. It causes objects to oppose any change in velocity. Newton's first law is often called the law of inertia.

A human body moving at 20 km h^{-1} has a lot of momentum. This momentum has to be lost every time the body and the go-kart stop moving. *Impulse* is the change in momentum that occurs when a force is applied over a period of time; that is, $\Delta \boldsymbol{p} = \boldsymbol{F} \times \Delta t$, where $\Delta \boldsymbol{p}$ is the change in momentum (kg m s^{-1}), $\boldsymbol{F}$ is the force (N) and Δt is the time period over which the force is acting (s). Many of the safety features required to ride the go-karts reduce the force component of the impulse, thus reducing the effect of an impact on the rider involved in a crash.

IMPULSE is the change in momentum of an object. It is also the product of the force and the time period over which the force acts.

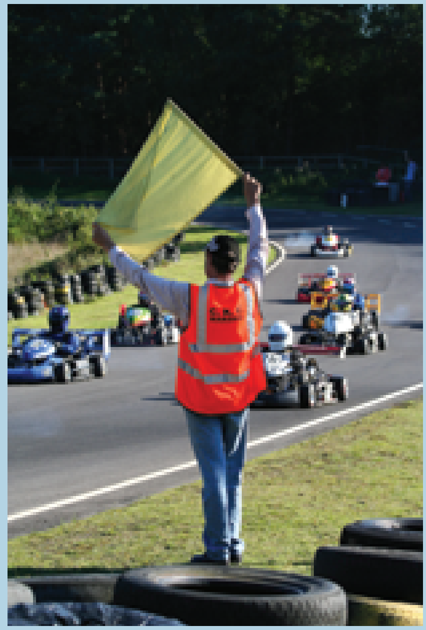


Figure fp.10 Operators of these rides have a duty to ensure that safe practices are followed.

Seat belts

A seat belt is one of the safety features incorporated into the design of go-karts. The first action of a seat belt is to stretch slightly to apply a stopping force to your body over a relatively long time period, much longer than would be done by the advertising board you have crashed into, for example. Increasing the time taken to stop your body means that a smaller force is required to achieve the same change in momentum. A smaller force means that less damage is done to you.

✓ Worked Example fp.4

An 80.0 kg person drives a go-kart into a tree at 20.0 km h^{-1}.

a Calculate the change in momentum of the person's body.

b Calculate the stopping force required if the time taken for the body's deceleration is:

  i 0.500 s (wearing a seat belt)

  ii 0.0100 s (not wearing a seat belt).

Solution

a $m = 80.0$ kg

$\boldsymbol{u} = 20.0$ km h^{-1} $= 5.56$ m s^{-1}

$\boldsymbol{v} = 0.00$ km h^{-1}

$$\Delta \boldsymbol{p} = \boldsymbol{p}_{final} - \boldsymbol{p}_{initial} = (m\boldsymbol{v} - m\boldsymbol{u})$$
$$= [(80.0)(0.00)] - [(80.0)(5.56)]$$
$$= -4.44 \times 10^2 \text{ kg m s}^{-1}$$

Hence, the change in momentum is 444 kg m s^{-1} in the opposite direction to the motion of the go-kart

b i $\Delta \boldsymbol{p} = -4.44 \times 10^2$ kg m s^{-1}

$\Delta t = 0.500$ s

$$\Delta \boldsymbol{p} = \boldsymbol{F}\Delta t$$

$$\boldsymbol{F} = \frac{\Delta \boldsymbol{p}}{\Delta t} = \frac{-4.44 \times 10^2}{0.500} = -8.89 \times 10^2 \text{ N}$$

So, the body experiences 889 N in the opposite direction to its motion, when wearing a seat belt.

b ii $\Delta \boldsymbol{p} = -4.44 \times 10^2$ kg m s^{-1}

$\Delta t = 0.0100$ s

$$\Delta \boldsymbol{p} = \boldsymbol{F}\Delta t$$

$$\boldsymbol{F} = \frac{\Delta \boldsymbol{p}}{\Delta t} = \frac{-4.44 \times 10^2}{0.0100} = -4.44 \times 10^4 \text{ N}$$

When the body is not wearing a seat belt, it experiences 44 400 N in the opposite direction of its motion (as it hits the tree).

Physics file

If the surface area of a seat belt increases, it will reduce the pressure on the person's body during the period of time that the force is acting, as pressure is equal to the force applied divided by the area over which the force is acting; therefore the larger the area the lower the pressure.

The second action of the seat belt is to spread the force over a relatively large contact area. This reduces the pressure experienced by the body. Lap-sash seat belts distribute the force over a much larger area than lap-only belts and, hence, are considered safer than lap-only belts. Many go-karts use harnesses that have two straps, one over each shoulder, to further increase the surface area in contact with the rider.

Turns

When you are in a go-kart car speeding around a corner, you feel as though you are being thrown towards the outside of the corner. Actually, you are experiencing something far more complicated, since there is no force acting to push you out. (If this answer surprises you, think about what is doing the pushing, and remember that 'moving' and 'inertia' are not forces.)

Let us look at a go-kart turning a corner. Just prior to the corner your inertia is straight ahead. The action of the car tyres on the road is to turn the car. Your body has mass and therefore has inertia, it tends to continue moving straight ahead, but now it is being forced to follow the path of the curve by the seat or side of the go-kart pushing on you. Because you can feel the pressure of the go-kart's seat and side pushing you sideways, it feels as if you are being thrown against the side of the go-kart. In actual fact, the go-kart is pushing against your side, forcing you to go around the corner!

The size of the force can be determined from the centripetal acceleration equation:

$$\boldsymbol{F}_c = \frac{mv^2}{\boldsymbol{r}}$$

where $\boldsymbol{F}_c$ is the centripetal force (N) and is always towards the centre of curvature, m is the mass of the object undergoing the circular motion (kg), v is the rotational or circular speed of the object (m s^{-1}) and $\boldsymbol{r}$ is the radius of curvature of the circular path.

Physics file

The term 'centrifugal force' is often used by mistake when people consider circular motion. When you sit next to the door in a car that goes around a corner, you feel as though you push against the door in a direction outwards from the centre of the curve. When you consider that for every action there is an equal and opposite reaction, then the centripetal force of the door on you is the action and the reaction force is you on the door, this force is the centrifugal force. But this force is not the one that causes you to turn in a circular path; the centripetal force is the one that does that!

Figure fp.11 A go-kart going around a corner at high speed

In this case, the size of the force experienced by a person will depend on that person's mass, the speed at which the car turns the corner and the radius of turning. Obviously, higher speeds and sharper curves result in greater accelerations and greater forces.

The centripetal force, experienced as a force exerted by the seat or side of the car on the driver, is mass-dependent; that is, a heavier person experiences a greater force. A greater force acting on a greater mass results in the same acceleration, so it is usual to discuss the acceleration (in 'g's) experienced by the riders as this is independent of mass.

Twist and turn, and shake it all about

The analysis of a complex or corkscrew manoeuvre on a roller-coaster can be quite intimidating. Remember that the resultant force experienced by the riders is merely the sum of all the forces acting at the time. Therefore, in a combination turn and loop, like a corkscrew or a tilted loop, the resultant force experienced is the vector sum of the circular motion effect caused by the horizontal turn, and the circular motion effect caused by the vertical loop. The only difficulty is that the analysis can involve three-dimensional resolution of all of the vectors if you choose the wrong frame of reference. Calculations are a little easier if you can choose a frame of reference that includes most of the forces and directions of travel.

Figure fp.12 Combinations of loops and turns provide even more excitement!

Physics file

Turning speeds are often expressed in revolutions per minute (rpm). The conversion is fairly simple. Ignoring the direction of travel, the velocity becomes a speed:

$$v = \frac{\text{distance}}{\text{time}} = \frac{2\pi r \times \text{rpm}}{60} = \text{m s}^{-1}$$

$$\text{rpm} = \frac{60 \times v}{2\pi r}$$

See 1.2 Circular motion in a horizontal plane on page 33.
See 1.3 Circular motion in a vertical plane on page 42.

Physics file

There are actually very few circular loop-the-loops used in roller-coasters. As there is a loss of kinetic energy as the coaster increases in height, there is a corresponding decrease in speed and therefore a decrease in the centripetal force. The centripetal force must remain large at the top so that we feel firm contact with the seat. Recall that centripetal force increases when the radius decreases, so if we can reduce the radius of the loop as the roller coaster travels up the loop, then the centripetal force can remain constant even though the speed is decreasing. On the way down, the radius increases so that the centripetal force is reduced as the coaster gains speed again. The loop that describes this shape is known as the clothoid loop.

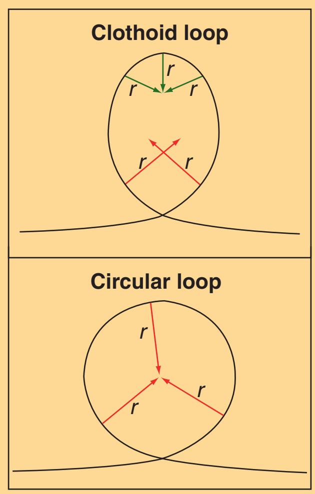


Figure fp.13 The changing radius of a clothoid loop enables a constant centripetal force despite a change in speed.

Exercises

E15 As a roller-coaster car reaches a turn, it is travelling at 6.00 m s^{-1}. If the designers of the ride want the horizontal centripetal acceleration felt by the passengers to be less than 2.00***g***, what is the minimum radius turn that could be placed at this location?

E16 To fit into the available space, a turn on a roller-coaster track has a radius of no more than 10.0 m. At what maximum speed can the car enter the turn if the acceleration felt by the riders is not to exceed 1.50***g***?

E17 A roller-coaster car with its maximum load of passengers weighs 1.00 tonne. If it travels at 15.0 m s^{-1} into a turn of radius 20.0 m, what force must the track be able to withstand?

E18 Because of track and roller-coaster car limitations, the designers cannot have a sideways acceleration of more than 1***g*** on a particular ride. How can the designers still create a ride in which the passengers experience 2.00***g***, without redesigning the car or track?

Figure fp.14 The first drop of 'Goliath'

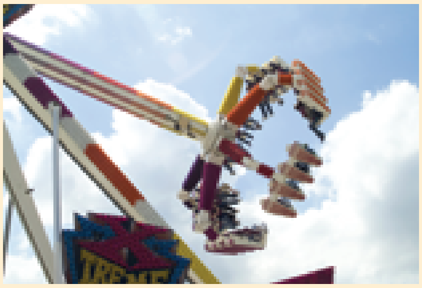


Figure fp.15 Some rides are just more complicated versions of others.

Spin and turn and up and down

Imagine sitting in a freely rotating bucket seat that is attached to a spinning platform that is attached to a larger spinning platform. This 'whirly-twirly' type of ride is quite common in fairgrounds; perhaps you have even ridden one. The motion you feel when on this ride is simple to analyse. First, you have a centripetal force and acceleration created by the turning of the large platform. The large platform usually turns at a constant rate, so this aspect of the sensation is constant.

The smaller platform also gives a centripetal force and acceleration. These are added to the force and acceleration provided by the larger platform, remembering that force and acceleration are both vectors, and must be added as vectors. At times, the two will be in the same direction, their magnitudes adding to each other, giving a maximum (which is the sum of the two), and at other times they will oppose each other, giving a minimum (which is the difference).

The times at which the two forces oppose or reinforce each other will depend on the periods of motion of each platform. Assuming constant rates of rotation for each platform, then these will be predictable and regular.

For even more excitement, the larger platform can be made to raise and lower, or tilt. This adds a vertical component to the sensation, and if the tilt is far enough, it adds a gravitational force component to the maximum and minimums mentioned before!

The analysis of the motion of this ride can be quite complex, so we will limit ourselves to calculating the maximum force and acceleration in the exercises.

Exercises

E19 In a whirly-twirly ride, the small platform has a radius of 2.00 m and rotates at 0.500 m s^{-1} and the large platform has a radius of 10.0 m and rotates at 2.00 m s^{-1}. What is the maximum centripetal force experienced by a 50.0 kg rider?

E20 In another whirly-twirly ride, the small platform has a radius of 2.00 m and rotates at 1.00 m s^{-1} and the large platform has a radius of 10.0 m and rotates at 0.600 m s^{-1}. What is the maximum acceleration experienced by the riders?

Bumper cars

Bumper cars are one of the 'old-school' rides; as such they are one of the rides that you are most likely to get your parents to go on. The physics involved in the bumper car ride is best explained using the concepts of momentum and the law of conservation of momentum.

See Momentum on page 438 on CD.
See Impulse on page 440 on CD.
See Conservation of energy on page 441 on CD.

CONSERVATION OF MOMENTUM: the momentum of a body is given by the mass times the velocity. In a system in which two or more objects interact, the sum of the momentum of the bodies before the interaction will equal the sum of the momentum of the bodies after the interaction.

Figure fp.16 When riding the bumper cars, rear-end collisions are inevitable!

In any collision, the law of conservation of momentum is the first consideration. Conservation of kinetic energy is not usually considered in motor-vehicle collisions, as these are usually not elastic collisions. This is because in these collisions most of the original kinetic energy of the cars changes into other forms of energy. Kinetic energy is used to change the shape of the cars and to produce heat and sound. Fortunately, a great deal of information about a collision can be found just from considering momentum.

One-dimensional collisions

First, we will consider a number of examples of collisions in one dimension. Later, this analysis will be extended to two dimensions by considering conservation of momentum independently in each perpendicular direction.

✓ Worked Example fp.5

A red bumper car (mass 808 kg) moving at 10.0 km h^{-1}, bumps into the back of an orange bumper car (801 kg) that is stationary on the track.
If the red bumper car stops as a result of the collision, what is the final speed of the orange bumper car?

Solution

$\boldsymbol{u}_{red} = 10.0 \text{ km h}^{-1} = 2.78 \text{ m s}^{-1}$

$m_{red} = 808$ kg

$\boldsymbol{v}_{red} = 0 \text{ m s}^{-1}$

$\boldsymbol{u}_{orange} = 0 \text{ m s}^{-1}$

$m_{orange} = 801$ kg

Momentum is conserved, so:

$$m_{red}\boldsymbol{u}_{red} + m_{orange}\boldsymbol{u}_{orange} = m_{red}\boldsymbol{v}_{red} + m_{orange}\boldsymbol{v}_{orange}$$

$$\boldsymbol{v}_{orange} = \frac{m_{red}\boldsymbol{u}_{red} + m_{orange}\boldsymbol{u}_{orange} - m_{red}\boldsymbol{v}_{red}}{m_{orange}}$$

$$= \frac{(808)(2.78) + (801)(0) - (808)(0)}{801}$$

$$= 2.80 \text{ m s}^{-1} \quad \text{or} \quad 1.01 \times 10^{1} \text{ km h}^{-1}$$

The speed of the orange bumper car is 2.80 m s^{-1} or about 10.1 km h^{-1} in the same direction as the red bumper car was initially travelling.

✓ Worked Example fp.6

A blue bumper car (mass 752 kg) moving at 25.0 km h^{-1} 'rear-ends' (i.e. runs into the back of) a red bumper car (808 kg) travelling at 15.0 km h^{-1} in the same direction as the blue car.
If the two vehicles stick together on collision, what will be the velocity of the wreckage?

Solution

$\boldsymbol{u}_{blue} = 25.0 \text{ km h}^{-1} = 6.94 \text{ m s}^{-1}$

$m_{blue} = 725 \text{ kg}$

$\boldsymbol{u}_{red} = 15.0 \text{ km h}^{-1} = 4.17 \text{ m s}^{-1}$

$m_{red} = 808 \text{ kg}$

$m_{red+blue} = 1533 \text{ kg}$

Momentum is conserved, so:

$$m_{blue}\boldsymbol{u}_{blue} + m_{red}\boldsymbol{u}_{red} = m_{red+blue}\boldsymbol{v}_{red+blue}$$

$$\boldsymbol{v}_{red+blue} = \frac{m_{blue}\boldsymbol{u}_{blue} + m_{red}\boldsymbol{u}_{red}}{m_{red+blue}}$$

$$= \frac{(725)(6.94) + (808)(4.17)}{1533}$$

$$= 5.48 \text{ m s}^{-1} \quad \text{or} \quad 1.97 \times 10^{1} \text{ km h}^{-1}$$

The speed of the combined bumper cars is 5.48 m s^{-1} or about 19.7 km h^{-1} in the same direction as the bumper cars were initially travelling.

According to police and insurance assessors, any vehicle that runs into the rear of another vehicle is automatically assumed to be in the wrong. This is one of the reasons why tailgating (i.e. driving very close to the rear of another car) is considered poor driving, but not on the bumper car track!

What happens if the vehicles are travelling towards each other when they collide?

Figure fp.17 While the fairground bumper car collisions are fun, real head-on collisions are considered to be the most serious type of crash.

Investigation

Use two motion detectors (e.g. Pasco) to investigate whether or not head-on collisions between dynamics trolleys obey the law of conservation of momentum.

In both the head-on and the rear-end collisions, the final velocity of the bumper cars will only apply for a short time. After a collision, the brakes of a vehicle will often lock on and the combination comes to a rapid stop.

Two-dimensional collisions

Very few motor-vehicle collisions occur in a straight line. Although analysis of collisions at an angle can be quite complex, the basic principles can be demonstrated by considering a simple example.

✓ Worked Example fp.7

If the blue and the red bumper cars in Worked Example fp.6 were travelling in opposite directions, they would have collided head-on.
Using the values given for initial speed and mass in Worked Example fp.6, determine the velocity of the combined bumper cars after a head-on collision.

Solution

Since the vehicles are moving in different directions, we will need to define one direction as being positive. For convenience, we will consider the direction of the red bumper car's motion as positive and the blue car's direction as negative. (The analysis works just as well if this is reversed, as long as you are consistent.) So:

$\overset{\longleftrightarrow}{- \quad +}$

$u_{blue} = -25.0 \text{ km h}^{-1} = -6.94 \text{ m s}^{-1}$

$m_{blue} = 725 \text{ kg}$

$u_{red} = +15.0 \text{ km h}^{-1} = +4.17 \text{ m s}^{-1}$

$m_{red} = 808 \text{ kg}$

$m_{red+blue} = 1533 \text{ kg}$

Momentum is conserved, so:

$$m_{blue}u_{blue} + m_{red}u_{red} = m_{red+blue}v_{red+blue}$$

$$v_{red+blue} = \frac{m_{blue}u_{blue} + m_{red}u_{red}}{m_{red+blue}}$$

$$= \frac{(725)(-6.94) + (808)(+4.17)}{1533}$$

$$= -1.09 \text{ m s}^{-1} \quad \text{or} \quad -3.92 \text{ km h}^{-1}$$

The speed of the combined bumper cars is -1.09 m s^{-1} or about -3.92 km h^{-1}. The negative sign indicates that they will be moving in the direction of the blue bumper car's original motion.

Figure fp.18 In real car crashes that occur at an angle, side curtain air bags would provide some level of protection against injury.

✓ Worked Example fp.8

What will be the velocity of the combined bumper cars if the red and the blue cars of Worked Example fp.6 collide at right angles to each other? For the sake of simplicity, assume that the velocities of the two bumper cars were at right angles to each other at the point of collision. The blue bumper car is travelling in the *x* direction and the red car is travelling in the *z* direction.

Solution

In this case, the law of conservation of momentum applies in each direction.
In the *x* direction:

– +

$\boldsymbol{u}_{blue} = 25.0\text{ km h}^{-1} = 6.94\text{ m s}^{-1}$

$m_{blue} = 725\text{ kg}$

Momentum in the *x* direction is conserved, so:

$$\boldsymbol{p}_{x\text{ final}} = m_{blue}\boldsymbol{u}_{blue} = (725)(6.94) = 5.03\times10^3\text{ kg m s}^{-1}$$

In the *z* direction:

– +

$\boldsymbol{u}_{red} = 15.0\text{ km h}^{-1} = 4.17\text{ m s}^{-1}$

$m_{red} = 808\text{ kg}$

Momentum in the *z* direction is conserved, so:

$$\boldsymbol{p}_{z\text{ final}} = m_{red}\boldsymbol{u}_{red} = (808)(4.17) = 3.37\times10^3\text{ kg m s}^{-1}$$

Adding the *x* and *z* vectors gives:

$\boldsymbol{p}_{x\text{ final}} = 5.03\times10^3\text{ kg m s}^{-1}$

$\boldsymbol{p}_{z\text{ final}} = 3.37\times10^3\text{ kg m s}^{-1}$

$$\boldsymbol{p}_{final}^2 = \boldsymbol{p}_{x\text{ final}}^2 + \boldsymbol{p}_{z\text{ final}}^2$$

$$\boldsymbol{p}_{final} = \sqrt{(5.03\times10^3)^2 + (3.37\times10^3)^2} = \sqrt{3.67\times10^7} = 6.05\times10^3\text{ kg m s}^{-1}$$

$$\tan\boldsymbol{p} = \frac{\boldsymbol{p}_{z\text{ final}}}{\boldsymbol{p}_{x\text{ final}}} = \frac{3.37\times10^3}{5.03\times10^3}$$

$$\boldsymbol{p} = \tan^{-1}(0.6689) = 33.8°$$

Thus is the magnitude and direction of the total final momentum is 6050 kg m s^{-1} at 33.8° from the motion of the blue bumper car. Since the wreckage of the two cars has a combined mass of 1533 kg, its final velocity can be found easily:

$\boldsymbol{p}_{final} = 6.05\times10^3\text{ kg m s}^{-1}$

$m = 1533\text{ kg}$

$$\boldsymbol{p}_{final} = m\boldsymbol{v}$$

$$\boldsymbol{v} = \frac{6.05\times10^3}{1533} = 3.95\text{ m s}^{-1}\text{ at } 33.8°\text{ from the path of the blue bumper car}$$

Therefore, the wreckage of the two vehicles will move at a speed of 3.95 m s^{-1} or 14.2 km h^{-1} at an angle of 33.8° to the direction of the blue bumper car's original motion.

Although this is a relatively simple example, the technique of applying conservation of momentum independently in perpendicular directions can be applied to much more complicated examples.

Investigation

1 Investigate the reasons for lowering the speed limit in suburban areas from 60 km h^{-1} to 50 km h^{-1}. Do you believe this change was justified?

2 Which is the safest car in Australia? Justify your answer.

Exercises

E21 A 420.0 kg yellow bumper car collides head-on with a 500.0 kg purple bumper car travelling in the opposite direction. Assuming the two vehicles stick together in the collision, calculate the velocity of the wreckage if the cars are travelling at the following speeds:

- **a** the yellow car at 13.0 km h^{-1} and the purple one at 7.00 km h^{-1}
- **b** 6.00 km h^{-1} each.

E22 A 410.0 kg orange bumper car travelling at 10.0 km h^{-1} collides head-on with a 350.0 kg silver bumper car. The two cars stick together in the collision and the combination does not move from the point of impact. Calculate the initial speed of the silver car.

E23 A 530.0 kg gold bumper car travelling at 15 km h^{-1} runs into the back of an 800.0 kg green car that is stopped on the track. Calculate the velocity of the cars after the collision, assuming that they stick together.

E24 A turquoise bumper car (weighing in at 425.0 kg) is travelling east at 6.00 km h^{-1}. It is hit by a white car (675.0 kg) heading south. If the combined cars end up sliding directly south-east, how fast was the white car going at the point of collision?

Experimental investigation

Design your favourite type of amusement park ride. Your design should take into consideration the maximum forces and accelerations experienced by the riders and contain a detailed account of what the riders will experience during the course of the ride.

Build a scale model of an existing theme park ride, or design a ride yourself and build a model of it to check your design. Your ride does not have to be built using model cars and people. For example, you could build a model roller-coaster from clear plastic tubing and marbles.

Investigation

1 Visit a theme park and take measurements to determine the force and acceleration experienced by a rider. Compare your results with those of your classmates to determine which ride results in the greatest force or acceleration experienced by riders.

2 Analyse an existing theme park ride in detail and calculate the forces and accelerations experienced by the riders. This might involve taking measurements of an existing ride during a visit to a theme park, or accessing the data another way. Your calculation of the forces and accelerations might be theoretical or practical. That is, you might take measurements of heights and slopes, and calculate the theoretical forces and accelerations, or you might take an accelerometer on the ride and measure the forces that way. If you are able to, do both and compare the theoretical results with the practical measurements.

1 Analysing motion

This (opposite) is not a picture of a junkyard. It is a freeway in California. Nearly 200 cars and trucks have collided, injuring dozens of people. There was thick fog at the time and so visibility was very poor. Evidently, motorists were driving too fast and too close for such conditions.

When you drive or travel in a car, you can experience what Sir Isaac Newton described in his laws of motion around 300 years ago. For a car to speed up or slow down, unbalanced forces must act on it. If the forces are balanced, the car continues in its original state of motion. If the car collides with another, they exert equal but opposite forces on each other. These forces are often extreme and can lead to drivers and passengers being injured or killed. In most parts of the world, the road toll is a serious issue. In Victoria, on average, about 350 people die as a result of car accidents each year.

A team from the Monash University Accident Research Centre is researching ways to make cars safer. The TAC SafeCar project is installing a variety of new devices and technologies into vehicles and testing their effectiveness in reducing accidents. One such device is 'intelligent speed adaptation' in which the speed limit is encoded into a digital map so that the car 'knows' what the speed limit is and informs the driver when this is exceeded. The 'forward collision warning system' sounds an alarm if the car is too close to the vehicle in front for the speed at which it is travelling. A 'reverse collision warning system' uses sonar to detect if objects behind the car are too close when the car is reversing. Should any alcohol vapour be detected in the car, the driver would need to blow into an in-car breathalyser to ensure that he or she is under the legal blood alcohol limit. If an accident occurs, an emergency response system in the car would automatically alert the ambulance service of the car's location. The research team hopes that positive trial results will encourage vehicle fleet owners and large organisations to implement some or all of these features into their cars.

Extra material that revisits and expands on 'Analysing motion' can be found on the *ePhysics* CD.

By the end of this chapter

you will have covered material from the study of analysing motion including:

- motion in one and two dimensions
- Newton's laws of motion
- projectile motion
- momentum and impulse
- conservation of momentum
- work, energy and conservation of energy
- elastic and inelastic collisions
- circular motion in horizontal planes and vertical planes
- banked corners.

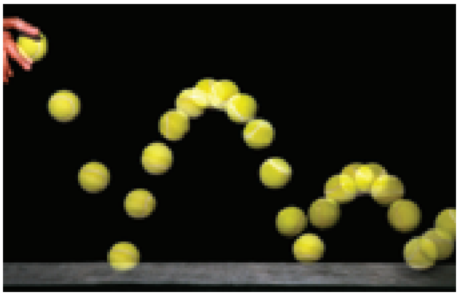


Figure 1.1

A multiflash photograph of a tennis ball that has been bounced on a hard surface. The ball moves in a parabolic path.

Practical activity

12 Projectile motion

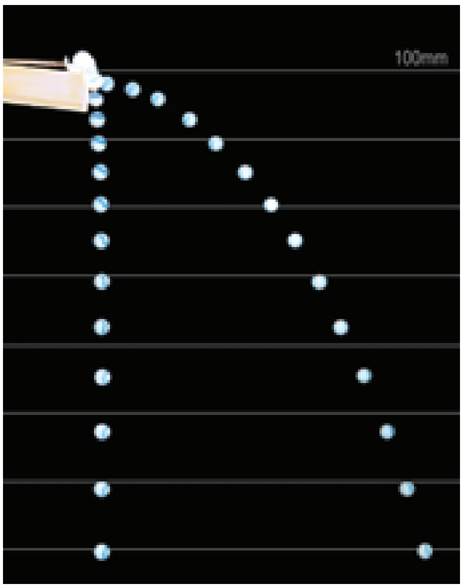


Figure 1.2

A multiflash photo of two golf balls released simultaneously. One ball was launched horizontally at 2.00 m s^{-1} while the other was released from rest. The projectile launched horizontally travels an equal horizontal distance during each flash interval, indicating that its horizontal velocity is constant. However, in the vertical direction, this projectile travels greater distances as it falls. In other words, it has a vertical acceleration. In fact, both balls have a vertical acceleration of 9.80 m s^{-2}, so they both fall at exactly the same rate and land at the same time. This shows that the two components of motion are independent: the horizontal motion of the launched projectile has no effect on its vertical motion (and vice versa).

1.1 Projectile motion

A projectile is any object that is thrown or projected into the air and is moving freely, i.e. it has no power source (such as a rocket engine) driving it. A netball as it is passed, a coin that is tossed and a gymnast performing a dismount are all examples of projectiles. If they are not launched vertically and if air resistance is ignored, projectiles move in *parabolic* paths.

> If air resistance is ignored, the only force acting on a ***projectile*** during its flight is its attraction to the Earth due to gravity, $\boldsymbol{F}_g$. This force is constant and always directed vertically downwards, and causes the projectile to continually deviate from a straight line path to follow a parabolic path.

- Given that the only force acting on a projectile is the force of gravity, $\boldsymbol{F}_g$, it follows that the projectile must have a vertical acceleration of 9.80 m s^{-2} downwards.
- The only force, $\boldsymbol{F}_g$, that is acting on a projectile is vertical and so it has no effect on the horizontal motion. The vertical and horizontal components of the motion are independent of each other and must be treated separately.
- There are no horizontal forces acting on the projectile, so the horizontal component of velocity will be constant.

> In the ***VERTICAL COMPONENT***, a projectile accelerates with the acceleration due to gravity, 9.80 m s^{-2} downward.
> In the ***HORIZONTAL COMPONENT***, a projectile moves with uniform velocity since there are no forces acting in this direction.

These points are fundamental to an understanding of projectile motion, and can be seen by studying Figure 1.2.

If *air resistance is ignored*, the motion of a projectile will be symmetrical around the vertical axis through the top of the flight. This symmetry extends to calculations involving speed, velocity and time as well as position. As seen in Figure 1.3, the projectile will take exactly the same time to reach the top of its flight as it will to travel from the top of its flight to the ground. At any given height, the speed of the projectile will be the same, and at any given height the velocities are related. On the way up, the angle for the velocity vector will be directed above the horizontal, whereas on the way down, the angle is the same but it is directed below the horizontal. For example, a projectile launched at 50 m s^{-1} at 80.0° above the horizontal will land at 50.0 m s^{-1} at 80.0° below the horizontal.

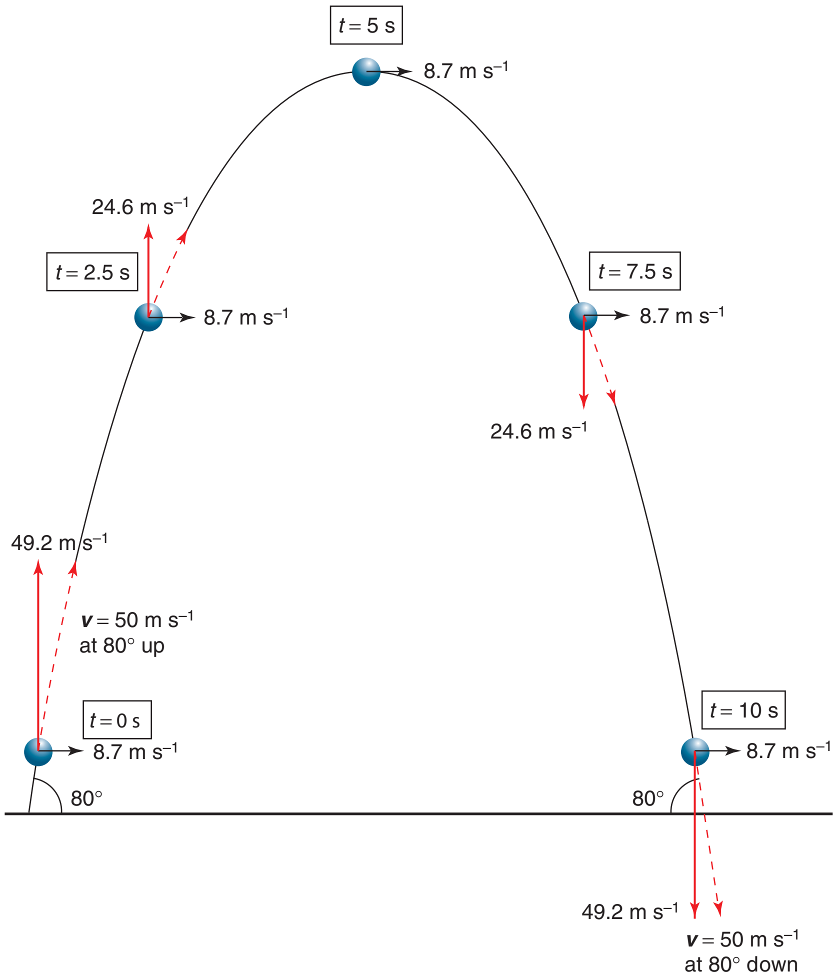


Figure 1.3
Ignoring air resistance, the horizontal velocity of the ball will remain the same, while the vertical component of the velocity will change with time. The motion of the projectile is symmetrical, and for a given height, the ball will have the same speed.

Tips for problems involving projectile motion

- Construct a diagram showing the motion and set the problem out clearly. Distinguish between the *x* and *y* dimension of the information supplied for each component of the motion.
- In the *horizontal dimension*, the velocity, $\boldsymbol{v}_x$, of the projectile is constant and so the only formula needed is $\boldsymbol{s}_x = \boldsymbol{v}_x \Delta t$.
- For the *vertical dimension*, the projectile is moving with a constant acceleration (9.80 m s^{-2} down), and so the equations of motion for uniform acceleration must be used.
- In the vertical dimension, it is important to clearly specify whether up or down is the positive or negative direction, and use this consistently throughout the problem. In this text we will use the convention that up is positive and down is negative, in accordance with the *y*-axis number line.

Physics file

A common misconception is that there is a driving force acting to keep a projectile moving through the air. This is a medieval understanding of motion. Such a force does not exist. For example, when you toss a ball across the room, your hand exerts a force on the ball as it is being thrown, but this force stops acting when the ball leaves your hand. There is no driving force propelling the ball along. Only the forces of gravity and air resistance act on the ball once it is in mid-air.

Physics file

It can be shown mathematically that the path of a projectile will be a parabola. Consider a projectile launched horizontally with velocity $\boldsymbol{v}_x$, and an acceleration down given by $\boldsymbol{g}$. Let $\boldsymbol{s}_x$ and $\boldsymbol{s}_y$ be the horizontal and vertical displacements respectively.
After a period of time Δt, the horizontal displacement is given by:

$$\boldsymbol{s}_x = \boldsymbol{v}_x \Delta t$$

$$\Delta t = \frac{\boldsymbol{s}_x}{\boldsymbol{v}_x} \qquad (1)$$

The vertical displacement after the same period of time Δt is:

$$\boldsymbol{s}_y = \boldsymbol{u}_y \Delta t + \tfrac{1}{2}\boldsymbol{g}\Delta t^2$$

As the initial velocity in the *y* direction is zero then:

$$\boldsymbol{s}_y = \tfrac{1}{2}\boldsymbol{g}\Delta t^2 \qquad (2)$$

Substituting (1) into (2) for the period of time, we get:

$$\boldsymbol{s}_y = \tfrac{1}{2}\boldsymbol{g}\left(\frac{\boldsymbol{s}_x}{\boldsymbol{v}_x}\right)^2$$

$$\boldsymbol{s}_y = \left(\frac{\boldsymbol{g}}{2\boldsymbol{v}_x^2}\right)\boldsymbol{s}_x^2$$

Since $\boldsymbol{g}$ and $\boldsymbol{v}_x$ are constant, then $\boldsymbol{s}_y \propto \boldsymbol{s}_x^2$, this is the relationship for a parabola.

Figure 1.4

Taller athletes have an advantage in the shot-put. They launch the shot from a greater height and so will achieve a greater distance. Recent female world champions have been around 2.0 m tall.

✓ Worked Example 1.1A

A golf ball of mass 64.0 g is hit horizontally from the top of a 40.0 m high cliff with a speed of 25.0 m s^{-1}. Assuming that the acceleration due to gravity is 9.80 m s^{-2} and ignoring air resistance, calculate the:

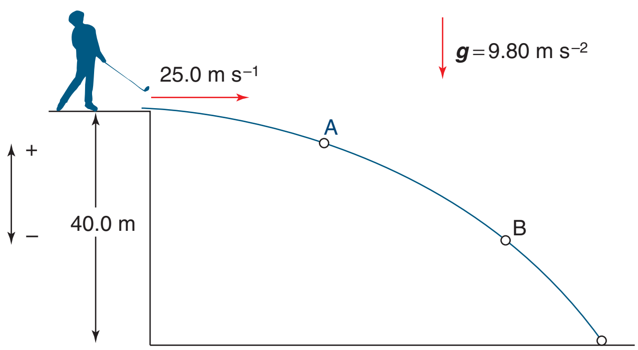
a time that the ball takes to land

b distance that the ball travels from the base of the cliff

c velocity of the ball as it lands

d net force acting on the ball at points A and B

e acceleration of the ball at points A and B.



Solution

a To find the time of flight of the ball, you need only consider the vertical component. The instant after it is hit, the ball is travelling only horizontally, so its initial vertical velocity is zero. Taking down as the negative direction:

$\updownarrow^{+}_{-}$

$\boldsymbol{u}_y = 0 \text{ m s}^{-1}$

$\boldsymbol{g} = -9.80 \text{ m s}^{-2}$

$\boldsymbol{s}_y = -40.0 \text{ m}$

$$\boldsymbol{s}_y = \boldsymbol{u}_y\Delta t + \tfrac{1}{2}\boldsymbol{g}\Delta t^2$$
$$= \tfrac{1}{2}\boldsymbol{g}\Delta t^2$$
$$\Delta t = \sqrt{\frac{2\boldsymbol{s}_y}{\boldsymbol{g}}} = \sqrt{\frac{2(-40.0)}{-9.80}}$$
$$= 2.86 \text{ s}$$

Hence, the ball takes 2.86 s to reach the ground.

b To find the horizontal distance travelled by the ball it is necessary to use the horizontal component of velocity.

$\boldsymbol{v}_x = 25.0 \text{ m s}^{-1}$

$\Delta t = 2.86 \text{ s}$

$$\boldsymbol{s}_x = \boldsymbol{v}_x\Delta t$$
$$= (25.0)(2.96)$$
$$= 71.4 \text{ m}$$

Hence, the ball lands 71.4 m from the base of the cliff (i.e. the range of the ball is 71.4 m).

c To determine the velocity of the ball as it lands, the horizontal and vertical components must be found separately and then added as vectors. From (a), the ball has been airborne for 2.86 s when it lands. The horizontal velocity of the ball is constant at 25.0 m s^{-1}. The vertical component of velocity when the ball lands is:

$\updownarrow^{+}_{-}$

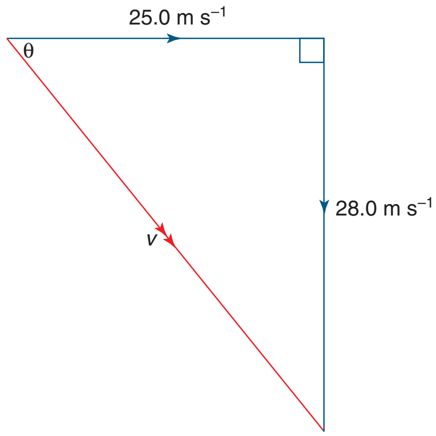
$\boldsymbol{u}_y = 0 \text{ m s}^{-1}$

$\boldsymbol{g} = -9.80 \text{ m s}^{-2}$

$\Delta t = 2.86 \text{ s}$

$$\boldsymbol{v}_y = \boldsymbol{u}_y + \boldsymbol{g}\Delta t$$
$$= 0 + (-9.80)(2.86)$$
$$= -28.0 \text{ m s}^{-1}$$

The actual velocity, **v**, of the ball is the vector sum of its vertical and horizontal components, as shown in the diagram. The magnitude of the velocity can be found by using Pythagoras's theorem:



$\boldsymbol{v}_x = 25.0 \text{ m s}^{-1}$

$\boldsymbol{v}_y = 28.0 \text{ m s}^{-1}$

$$\boldsymbol{v}^2 = \boldsymbol{v}_x^{\,2} + \boldsymbol{v}_y^{\,2}$$
$$= \sqrt{(25.0)^2 + (28.0)^2}$$
$$= 37.5 \text{ m s}^{-1}$$
$$\tan\theta = \frac{\boldsymbol{v}_y}{\boldsymbol{v}_x}$$
$$\theta = \tan^{-1}\frac{28.0}{25.0}$$
$$= 48.2°$$

When the ball hits the ground, it has a velocity of 37.5 m s^{-1} at an angle of 48.2° below the horizontal.

d If air resistance is ignored, the only force acting on the ball throughout its flight is its weight. Therefore the net force that is acting at point A and point B (and everywhere else!) is:

$m = 64.0 \times 10^{-3}$ kg

$\boldsymbol{g} = -9.80 \text{ m s}^{-2}$

$$\boldsymbol{F}_{\text{net}} = m\boldsymbol{g}$$
$$= (64.0 \times 10^{-3})(-9.80)$$
$$= -6.27 \times 10^{-1} \text{ N}$$

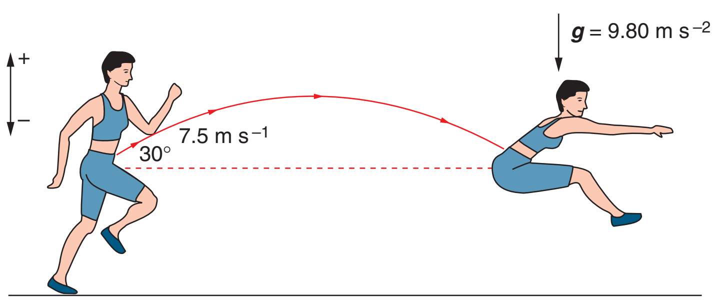
Hence, the net force on the ball is 0.627 N downwards.

e Since the ball is in free-fall, the acceleration of the ball at all points is equal to that determined by gravity, i.e. 9.80 m s^{-2} down.

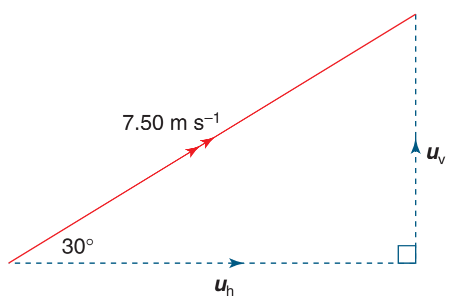
✓ Worked Example 1.1B

A 65.0 kg athlete in a long-jump event leaps with a velocity of 7.50 m s^{-1} at 30.0° to the horizontal. Treating the athlete as a point mass, ignoring air resistance, and using $\boldsymbol{g}$ as −9.80 m s^{-2}, calculate the:

a horizontal component of the initial velocity
b vertical component of the initial velocity
c velocity when at the highest point
d maximum height gained by the athlete
e total time for which the athlete is in the air
f horizontal distance travelled by the athlete's centre of mass (assuming that it returns to its original height)
g athlete's acceleration at the highest point of the jump.



Solution



In this problem, the upward direction will be taken as positive. The horizontal and vertical components of the initial velocity can be found by using trigonometry.

a As shown in the vector diagram, the horizontal component, $\boldsymbol{v}_x$, of the athlete's initial velocity is:

$\boldsymbol{v} = 7.50 \text{ m s}^{-1}$

$$\boldsymbol{v}_x = \boldsymbol{v} \times \cos 30.0°$$
$$= (7.50)\cos 30.0°$$
$$= 6.50 \text{ m s}^{-1}$$

This horizontal velocity of 6.50 m s^{-1} remains constant throughout the jump.

b Again referring to the diagram, the vertical component, $\boldsymbol{u}_y$, of the initial velocity of the athlete is:

$$\boldsymbol{v} = 7.50 \text{ m s}^{-1} \qquad \boldsymbol{u}_y = \boldsymbol{v} \times \sin 30.0°$$
$$= (7.50)\sin 30.0°$$
$$= 3.75 \text{ m s}^{-1}$$

Hence, the initial velocity of the athlete is 3.75 m s^{-1} upwards.

c At the highest point, the athlete is moving horizontally. The vertical component of the velocity at this point is therefore zero. The actual velocity is given by the horizontal component of the velocity throughout the jump. This was found in part a to be 6.50 m s^{-1} in the horizontal direction.

d To find the maximum height that is gained, we must work with the vertical component. As explained in part c, at the maximum height the athlete is moving horizontally and so the vertical component of velocity at this point is zero. The vertical displacement of the athlete to the highest point is the maximum height that was reached:

$$\updownarrow^{+}_{-} \qquad \boldsymbol{v}_y^2 = \boldsymbol{u}_y^2 + 2\boldsymbol{g}\boldsymbol{s}_y$$
$$\boldsymbol{u}_y = 3.75 \text{ m s}^{-1} \qquad \boldsymbol{s}_y = \frac{\boldsymbol{v}_y^2 - \boldsymbol{u}_y^2}{2\boldsymbol{g}}$$
$$\boldsymbol{g} = -9.80 \text{ m s}^{-2} \qquad = \frac{(0)^2 - (3.75)^2}{2(-9.80)}$$
$$\boldsymbol{v}_y = 0 \text{ m s}^{-1} \qquad = 0.717 \text{ m}$$

i.e. the centre of mass of the athlete rises by a maximum height of 71.7 cm.

e The motion is symmetrical as the athlete takes off and lands at the same height, so the time to complete the jump will be double that taken to reach the maximum height. First, the time to reach the highest point must be found. Using the vertical component:

$$\updownarrow^{+}_{-} \qquad \boldsymbol{v}_y = \boldsymbol{u}_y + \boldsymbol{g}\Delta t$$
$$\boldsymbol{u}_y = 3.75 \text{ m s}^{-1} \qquad \Delta t = \frac{\boldsymbol{v}_y - \boldsymbol{u}_y}{\boldsymbol{g}}$$
$$\boldsymbol{g} = -9.80 \text{ m s}^{-2} \qquad \Delta t = \frac{0 - 3.75}{-9.80}$$
$$\boldsymbol{v}_y = 0 \text{ m s}^{-1} \qquad = 0.383 \text{ s}$$

The time for the complete flight is double the time to reach maximum height, i.e. total time in the air: $\Delta t_{total} = 2 \times 0.3826 = 0.765$ s. Note that although Δt was 0.383, double that is 0.765 as the exact number in the calculator was doubled, not the rounded off number written on the page.

f To find the horizontal distance for the jump, we must work with the horizontal component. From part e, the athlete was in the air for a time of 0.765 s and so:

$$\boldsymbol{v}_x = 6.50 \text{ m s}^{-1} \qquad \boldsymbol{s}_x = \boldsymbol{v}_x \Delta t$$
$$\Delta t = 0.765 \text{ s} \qquad = (6.50)(0.765)$$
$$= 4.97 \text{ m}$$

i.e. the athlete jumps a horizontal distance of 4.97 m.

g At the highest point of the motion, the only force acting on the athlete is that due to gravity. The acceleration will therefore be 9.80 m s^{-2} down.

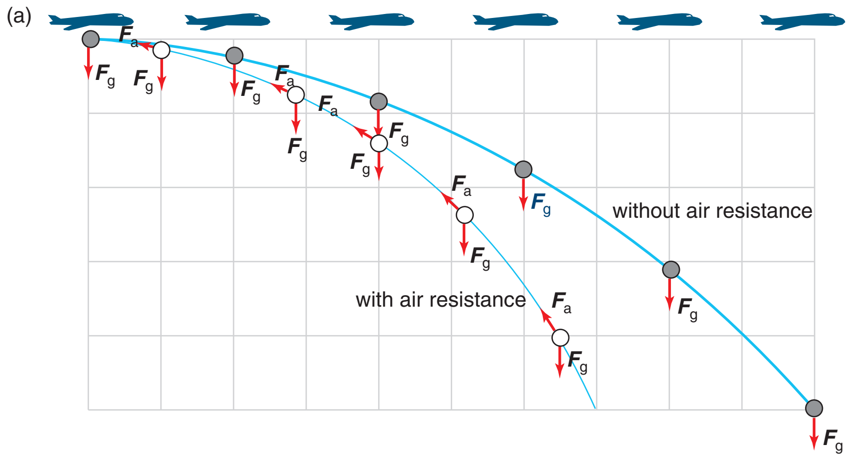
Physics file

A conservation of energy approach can also be used for solving projectiles problems. When air resistance can be ignored, the sum of the gravitational potential energy and kinetic energy of the projectile (i.e. its *mechanical* energy) is the same at all points in its flight. At the lowest point in its flight, gravitational potential energy is a minimum and kinetic energy is a maximum. At the highest point, the opposite occurs. The mass needs to be given to determine the actual energy values, but is not required to find the other properties such as acceleration, speed and displacement.

The effect of air resistance

In throwing events such as the javelin and discus, new records are not accepted if the wind is providing too much assistance to the projectile. In football games, kicking with the wind is generally an advantage to a team; and in cricket, bowling with the wind, across the wind or against the wind can have very different effects on the flight of the ball. The interaction between a projectile and the air can have a significant effect on the motion of the projectile, particularly if the projectile has a large surface area and a relatively low mass.

Figure 1.5 shows a food parcel being dropped from a plane moving at a constant velocity. If air resistance is ignored, the parcel falls in a parabolic arc. It would continue moving horizontally at the same rate as the plane; that is, as the parcel falls it would stay directly beneath the plane until it hits the ground. The effect of air resistance is also shown. Air resistance (or drag) is a retarding force and it acts in a direction that is opposite to the motion of the projectile. If air resistance is taken into account, there are now two forces acting: weight, $\boldsymbol{F}_g$, and air resistance, $\boldsymbol{F}_{air}$. Therefore, the sum of the forces, $\Sigma\boldsymbol{F}$, that acts on the projectile is *not* vertically down. The magnitude of the air resistance force is greater when the speed of the body is greater.



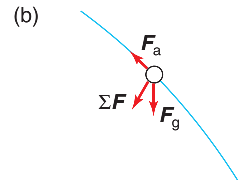


Figure 1.5
(a) The path of a food parcel dropped from a plane. If the plane maintains a constant speed after dropping the parcel and assuming no air resistance, the parcel will fall in a parabolic path and remain directly below the plane. Air resistance makes the parcel fall more slowly, over a shorter path. (b) When air resistance is acting, the net force on the parcel is not vertically down.

Physics in action — Why ballet dancers seem to float in the air

The grand jeté is a ballet movement in which dancers leap across the stage and appear to float in the air for a period of time. They position their arms and legs to give the impression that they are floating gracefully through the air. The dancer's centre of mass follows a parabolic path. Once the dancer is in mid-air, there is nothing that he or she can do to alter this path. However, by raising their arms and legs, dancers can raise the position of their centre of mass so that it is higher in the torso. The effect of this is that the dancer's head follows a lower and flatter line than it would have taken if the limbs had not been raised during the leap. The smoother and flatter line taken by the head gives the audience the impression of a graceful floating movement across the stage.

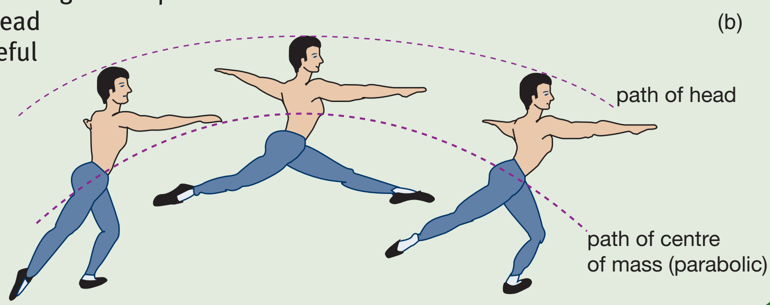


Figure 1.6
(a) Performing the grand jeté. (b) As the position of the centre of mass moves higher in the body, the head of the dancer follows a flatter path and this gives the audience the impression of a graceful floating movement.

1.1 SUMMARY Projectile motion

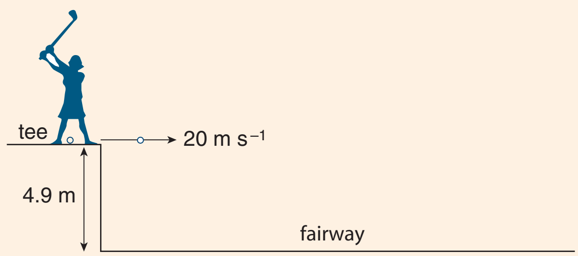
- Projectiles move in parabolic paths that can be analysed by considering the horizontal and vertical components of the motion.
- If air resistance is ignored, the only force acting on a projectile is the force of gravity, $\boldsymbol{F}_g$. This results in the projectile having a vertical acceleration of 9.80 m s^{-2} down during its flight.
- The horizontal velocity $\boldsymbol{v}_x$ of a projectile remains constant throughout its flight if air resistance is ignored.
- An object initially moving horizontally, but free to fall, will fall at exactly the same rate, and in the same time, as an object falling vertically from the same height.
- At the point of maximum height, a projectile is moving horizontally. Its velocity at this point is given by the horizontal component of its velocity $\boldsymbol{v}_x$ as the vertical component $\boldsymbol{v}_y$ equals zero.
- When air resistance is significant, the net force acting on a projectile will not be vertically down, nor will its acceleration. Under these conditions, the path of the projectile is not parabolic.

1.1 Questions

For the following questions, assume that the acceleration due to gravity is 9.80 m s^{-2} and ignore the effects of air resistance unless otherwise stated.

1 A golfer practising on a range with an elevated tee 4.90 m above the fairway is able to strike a ball so that it leaves the club with a horizontal velocity of 20.0 m s^{-1}.

- **a** How long after the ball leaves the club will it land on the fairway?
- **b** What horizontal distance will the ball travel before striking the fairway?
- **c** What is the acceleration of the ball 0.500 s after being hit?
- **d** Calculate the speed of the ball 0.800 s after it leaves the club.
- **e** With what speed will the ball strike the ground?

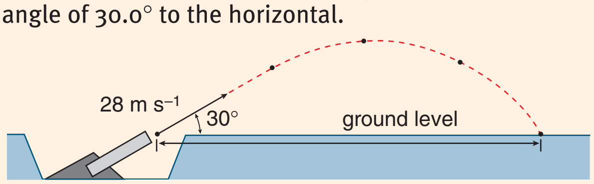


2 A bowling ball of mass 7.50 kg travelling at 10.0 m s^{-1} rolls off a horizontal table 1.00 m high.

- **a** Calculate the ball's horizontal velocity just as it strikes the floor.
- **b** What is the vertical velocity of the ball as it strikes the floor?
- **c** Calculate the velocity of the ball as it reaches the floor.
- **d** What time interval has elapsed between the ball leaving the table and striking the floor?
- **e** Calculate the horizontal distance travelled by the ball as it falls.
- **f** Draw a diagram showing the forces acting on the ball as it falls towards the floor.

The following information applies to questions 3–8.

A physics class researching projectile motion constructs a device that can launch a cricket ball. The launching device is designed so that the ball can be launched at ground level with an initial velocity of 28.0 m s^{-1} at an angle of 30.0° to the horizontal.



3 Calculate the horizontal component of the velocity of the ball:

- **a** initially
- **b** after 1.00 s
- **c** after 2.00 s.

4 Calculate the vertical component of the velocity of the ball:

- **a** initially
- **b** after 1.00 s
- **c** after 2.00 s.

5
- **a** At what time will the ball reach its maximum height?
- **b** What is the maximum height achieved by the ball?
- **c** What is the acceleration of the ball at its maximum height?

6
- **a** At which point in its flight will the ball experience its minimum speed?
- **b** What is the ball's minimum speed during its flight?
- **c** At what time does this minimum speed occur?
- **d** Draw a diagram showing the forces acting on the ball at the maximum height.

7
- **a** At what time after being launched will the ball return to the ground?
- **b** What is the velocity of the ball as it strikes the ground?
- **c** Calculate the horizontal range of the ball.

8 If the effects of air resistance were taken into account, which one of the following statements would be correct?

- **A** The ball would have travelled a greater horizontal distance before striking the ground.
- **B** The ball would have reached a greater maximum height.
- **C** The ball's horizontal velocity would have been continually decreasing.

9 A softball of mass 250.0 g is thrown with an initial velocity of 16.0 m s^{-1} at an angle θ to the horizontal. When the ball reaches its maximum height, its kinetic energy is 16.0 J.

- **a** What is the maximum height achieved by the ball from its point of release?
- **b** Calculate the initial vertical velocity of the ball.
- **c** What is the value of θ?
- **d** What is the speed of the ball after 1.00 s?
- **e** What is the displacement of the ball after 1.00 s?
- **f** How long after the ball is thrown will it return to the ground?
- **g** Calculate the horizontal distance that the ball will travel during its flight.

10 During training, an aerial skier takes off from a ramp that is inclined at 40.0° to the horizontal and lands in a pool that is 10.0 m below the end of the ramp. If she takes 1.50 s to reach the highest point of her trajectory, calculate the:

- **a** speed at which she leaves the ramp
- **b** maximum height above the end of the ramp that she reaches
- **c** time for which she is in mid-air.

1.2 Circular motion in a horizontal plane

Circular motion is common throughout the Universe. On the smallest scale, electrons travel around atomic nuclei in circular paths; on a bigger scale, the planets orbit the Sun in roughly circular paths; and on an even grander scale, stars can travel in circular paths around the centre of their galaxies. In this section we will be introducing circular motion in a horizontal plane.

Uniform circular motion

An athlete in a hammer throw event is swinging the hammer in a horizontal circle with a constant speed of 25.0 m s^{-1} (Figure 1.10). As the hammer travels in its circular path, its speed is constant but its velocity is continually changing. (Velocity is a vector, so the direction of motion is significant.) The velocity of the hammer at any instant is tangential to its path. At one instant the hammer is travelling at 25.0 m s^{-1} north, then an instant later at 25.0 m s^{-1} west, then 25.0 m s^{-1} south, and so on.

Let us say that an object is moving with a constant speed v in a circle of radius r metres and takes T seconds to complete one revolution. This time taken to move once around the circle is called the period of the motion. In completing one circle, the distance travelled by the object is equal to the circumference of the circle, $C = 2\pi r$. This can be used to find the average speed of the object undergoing the circular motion.

The ***AVERAGE SPEED*** of an object moving in a circular path is:

$$v = \frac{\text{distance}}{\text{time period}} = \frac{2\pi r}{T}$$

where v is the circular speed in metres per second (m s^{-1}), r is the radius of circular path of the object in metres (m) and T is the period of time taken for the object to go around the circular path once, in seconds (s).

✓ Worked Example 1.2A

An athlete is swinging a hammer of mass 7.00 kg in a circular path of radius 1.50 m. Calculate the speed of the hammer if it completes 3.00 revolutions per second.

Solution

$r = 1.50$ m

$f = 3.00\ s^{-1}$

$$T = \frac{1}{f} = \frac{1}{3.00}$$

$$T = 0.333\ s$$

$$v = \frac{2\pi r}{T} = \frac{2\pi(1.50)}{0.333}$$

$$= 28.3\ m\ s^{-1}$$

Hence the circular speed is 28.3 m s^{-1}. Note that this is the scalar quantity speed v, not the vector quantity velocity $\mathbf{v}$.

Physics file

These wind generators are part of a wind farm at Walkaway, 25 km south of Geraldton. Each tower is 50 m high, as high as one of the WACA light towers. Each blade is 29 m long and they rotate a constant 19 revolutions per minute. From this information, you should be able to calculate that the tip of each blade is travelling at around 220 km h^{-1}!

Figure 1.7
The tips of these wind-generator blades are travelling in circular paths at speeds of over 200 km h^{-1}.

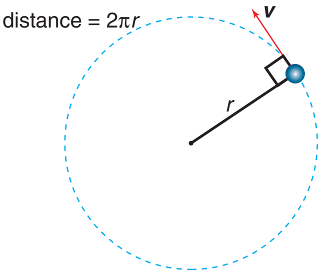


Figure 1.8
The average speed of an object moving in a circular path is given by the distance travelled in one revolution (the circumference) divided by the time taken (the period).

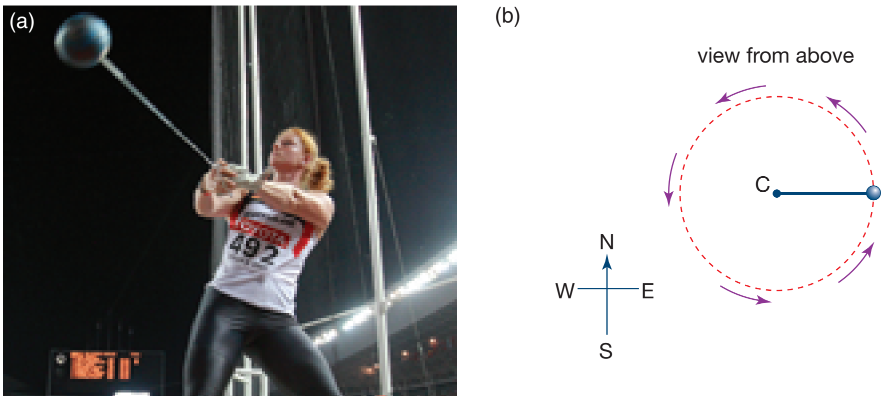


Figure 1.9

(a) The hammer, a ball of steel of mass 7.26 kg for men and 4.0 kg for women, is swung in a circular path before being released. The world record distance for a hammer throw is almost 90 m for men and almost 80 m for women.
(b) The velocity of any object moving in a uniform circular path is continually changing, even though its speed remains constant.

Physics file

The track on a CD starts at the centre and works its way to the outside of the disk. Compact disks spin at varying rates, depending on the radius of the track that is being played. When the inner track is being played, the disk rotates at about 500 revolutions per minute (rpm). This is 8.30 revolutions each second.

Practical activity

13 Centripetal force

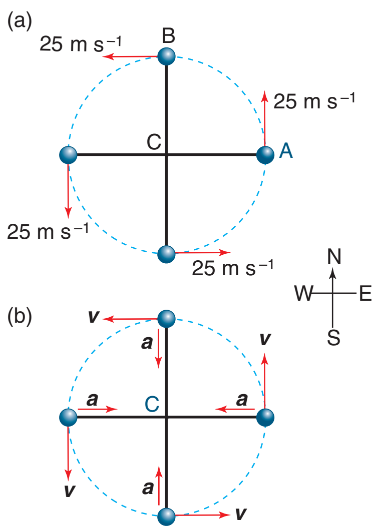


Figure 1.10

(a) The velocity of the hammer at any instant is tangential to its path.
(b) A body moving in a circular path has an acceleration towards the centre of the circle. This is known as a centripetal acceleration.

✓ Worked Example 1.2B

A compact disk is rotating at a rate of 8.00 revolutions per second. Calculate the speed of a point on the inner track of the CD which is moving in a circular path of radius 2.25 cm.

Solution

If the disk is rotating at 8.00 revolutions per second, its period is:

$f = 8.00\ \text{s}^{-1}$

$$T = \frac{1}{f} = \frac{1}{8.00} = 0.125\ \text{s}$$

The speed of a point on the inner track is:

$r = 2.25 \times 10^{-2}\ \text{m}$

$$v = \frac{2\pi r}{T} = \frac{2\pi(2.25 \times 10^{-2})}{0.125} = 1.13\ \text{m s}^{-1}$$

At the outer rim of the disk, it has slowed to around 200 rpm or 3.30 revolutions per second. This is done to ensure that, as the disk is played, the laser beam can be drawn along the track at a constant rate.

Centripetal acceleration

When an object moves in a circular path, its *velocity is changing*. In Figure 1.10a, a hammer is shown at various points as it travels in a circular path. When it is at point A, the hammer is moving north at 25 m s^{-1}. When at point B, it is moving west at 25 m s^{-1}, and so on. Therefore, the hammer is *accelerating* even though its speed is not changing. This acceleration is known as *centripetal acceleration*. (Centripetal means 'centre-seeking', which should be a reminder of its direction.) The hammer is continually deviating inwards from its straight-line direction and so has an *acceleration towards the centre*. However, even though the hammer is accelerating towards the centre of the circle, it never gets any closer to the centre (Figure 1.10b).

CENTRIPETAL ACCELERATION is always directed towards the centre of the circular path and is given by:

$$\boldsymbol{a}_c = \frac{v^2}{r}$$

A substitution can be made for the speed of the object in this equation.

$$\boldsymbol{a}_c = \frac{v^2}{r} = \frac{\left(\frac{2\pi r}{T}\right)^2}{r} = \frac{4\pi^2 r}{T^2}$$

where $\boldsymbol{a}_c$ is the centripetal acceleration in metres per second per second (m s^{-2}), v is the circular speed in metres per second (m s^{-1}), r is the radius of the circle in metres (m) and T is the period of rotation in seconds (s). Note that the centripetal acceleration is a vector, its direction is towards the centre of rotation.

Forces that cause circular motion

As with all forms of motion, an analysis of the forces that are acting is needed if we are to understand why circular motion occurs. In the hammer throw we have been looking at, the ball is continually accelerating, so it follows from Newton's second law that there must be an *unbalanced force* continuously acting on it. The unbalanced force that gives the hammer ball its acceleration towards the centre of the circle is known as a *centripetal force*. In every case where there is circular motion, a *real force* is necessary to provide the centripetal force (Figure 1.12).

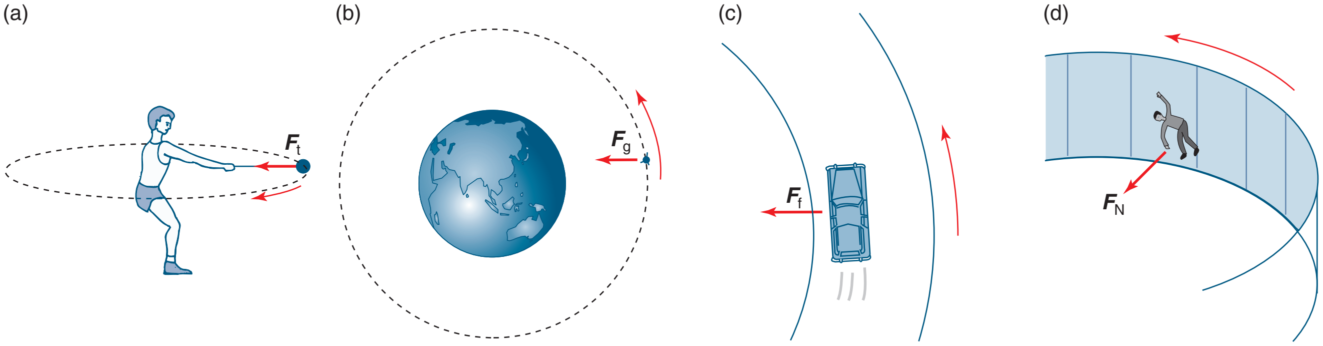


Figure 1.12
The centripetal force that produces a centripetal acceleration and hence a circular motion is provided by different real forces. (a) In a hammer throw or for any other object rotated while attached to an arm or wire, it is the tension in the arm or wire that provides the centripetal force. (b) For planets and satellites, the gravitational attraction to the central body provides the centripetal force. (c) For a car on a roundabout, it is the friction between the tyres and the road. (d) For a person in the Gravitron it is the normal force from the wall. Although the person feels that they are being pinned to the wall, the wall is in fact applying a force to their body.

Now, consider the consequences if the unbalanced force ceases to act. In the example of the hammer thrower, if the tension in the wire became zero because the thrower released the ball, there is no force causing the ball to change direction, so it then moves in a straight line tangential to its circular path (Figure 1.13), as would be expected from Newton's first law.

The centripetal force that causes centripetal acceleration for an object of mass m can be calculated from Newton's second law:

CENTRIPETAL FORCE is given by:

$$\boldsymbol{F}_c = m\boldsymbol{a}_c = \frac{mv^2}{r} = \frac{m4\pi^2 r}{T^2}$$

where $\boldsymbol{F}_c$ is the accelerating centripetal force on the object in newtons (N), m is the mass in kilograms (kg), v is the circular speed in metres per second (m s^{-1}), r is the radius of the circular path in metres (m) and T is the period of rotation in seconds (s). Note that centripetal force is also a vector with its direction towards the centre of the circle.

Physics file

People moving in circular paths often mistakenly think that there is an outwards force acting on them. For example, riders on the Gravitron will 'feel' a force pushing them into the wall. This outwards force is commonly known as a centrifugal (meaning 'centre-fleeing') force. This force does not actually exist in an inertial frame of reference. The riders think that it does because they are in the rotating frame of reference. From outside the Gravitron, it is evident that there is a force from the wall that is causing them to follow a circular path, and that in the absence of this force they would not 'fly outwards' but move at a tangent to their circle (see Figure 1.11).

Figure 1.11
There is a large force from the wall (a normal force) that causes these people to travel in a circular path.

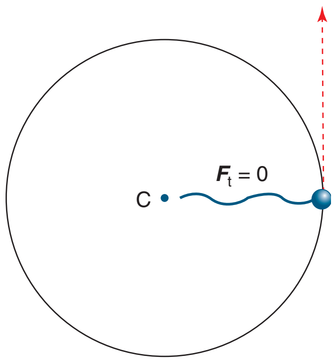


Figure 1.13
When a hammer ball is released by the thrower, it will travel in a straight line at a tangent to its circular path.

The force needed to make a body travel in a circular path (the centripetal force) will therefore increase if the mass of the body is increased, the speed of the body is increased, or the period of the motion is decreased.

The *kinetic energy* of a body that is moving horizontally with *uniform circular motion* remains *constant* even though an unbalanced force is continually acting on it. This is because the force is always *perpendicular* to the motion of the body. As was discussed earlier in this chapter, $W_d = \boldsymbol{F}_x \boldsymbol{s}_x$, but the force is directed at 90° to the direction of the instantaneous displacement, so there is no component of the displacement in the direction of the force. Therefore, the centripetal force does not do any work on the body. The force does, however, change the direction of motion of the body, while its speed remains constant.

✓ Worked Example 1.2C

An athlete in a hammer throw event is swinging the ball of mass 7.00 kg in a horizontal circular path. Calculate the tension in the wire if the ball is:

a moving at 20.0 m s^{-1} in a circle of radius 1.60 m

b moving at 25.0 m s^{-1} in a circle of radius 1.20 m.

Solution

a The centripetal acceleration is:

$v = 20.0$ m s^{-1} $\quad$ $r = 1.60$ m

$$a_c = \frac{v^2}{r} = \frac{(20.0)^2}{1.60} = 2.50 \times 10^2 \text{ m s}^{-2}$$
towards the centre

The tension is producing circular motion:

$m = 7.00$ kg $\quad$ $a_c = 2.50 \times 10^2$ m s^{-2}

$$F_c = ma_c = (7.00)(2.50 \times 10^2) = 1.75 \times 10^3 \text{ N}$$
towards the centre

Hence, the centripetal force is 1.75×10^3 N towards the centre.

b $v = 25.0$ m s^{-1} $\quad$ $r = 1.20$ m

$$a_c = \frac{v^2}{r} = \frac{(25.0)^2}{1.20} = 5.20 \times 10^2 \text{ m s}^{-2}$$
towards the centre

The tension is producing circular motion:

$m = 7.00$ kg $\quad$ $a_c = 5.20 \times 10^2$ m s^{-2}

$$F_c = ma_c = (7.00)(5.20 \times 10^2) = 3.66 \times 10^3 \text{ N}$$
towards the centre

Hence, the centripetal force is 3.66×10^3 N towards the centre.

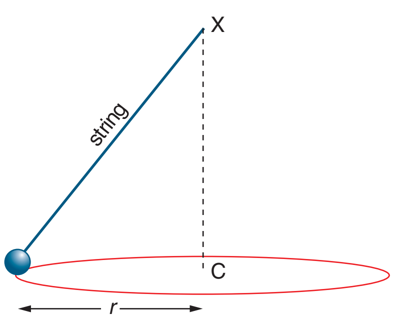


Figure 1.14
This yo-yo is travelling in a horizontal circular path of radius *r*. The centre of its circular motion is at C.

Ball on a string

You might have at one time played around with a yo-yo and swung it over your head in a *horizontal* circle. If you were swinging the yo-yo slowly, the string would swing down at a large angle close to your body. If you swung the yo-yo faster, the string would become much closer to horizontal. In fact, it is not possible for the string to be horizontal, although as the speed increases, the closer to horizontal it becomes. This system is known as a conical pendulum.

Consider the ball shown in Figure 1.14. It is attached to string, but the centre of its circular path is not the end of the string at X. It is at C, as shown on the diagram. Similarly, the radius, *r*, of its path is not the same as the length of the string. If an angle is known, trigonometry can be used to find this.

Banked corners

Cars and bikes can travel much faster around corners when the road or track surface is inclined or *banked* at some angle to the horizontal. Banking is most obviously used at cycling velodromes or motor sport events such as NASCAR races. Road engineers also design roads to be banked in places where there are sharp corners such as exit ramps on freeways.

When cars travel in circular paths on horizontal roads, they are relying on the force of friction between the tyres and the road. Friction provides the sideways force that makes the car turn.

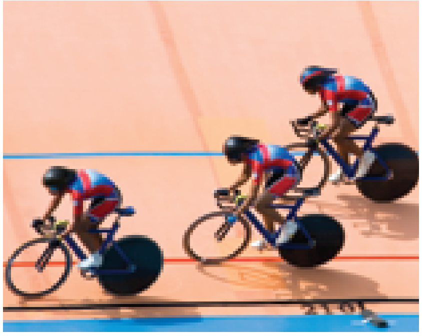
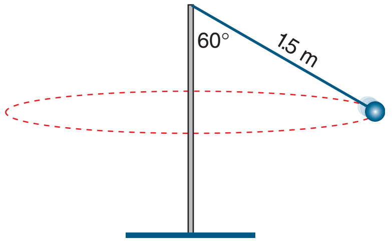


Figure 1.15
The cyclists on this banked velodrome track are cornering at speeds far higher than they could use on a flat track. The cyclists on the velodrome do not need to rely on friction to turn and experience a larger normal force than usual.

✓ Worked Example 1.2D

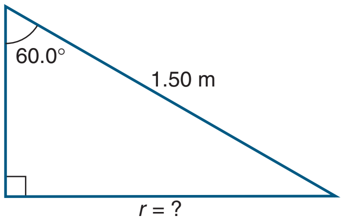
During a game of Totem-Tennis, the ball of mass 57.0 g is swinging freely in a horizontal circular path. The cord is 1.50 m long and is at an angle of 60.0° to the vertical shown in the diagram.



a Calculate the radius of the ball's circular path.
b Draw and identify the forces that are acting on the ball at the instant shown in the diagram.
c Determine the net force that is acting on the ball at this time.
d Calculate the size of the tensile force in the cord.
e How fast is the ball travelling at this time?

Solution

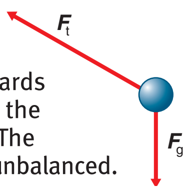
a The centre of the circular path is not at the top end of the cord, but is at the point where the pole cuts the horizontal plane of the path of the ball. Trigonometry and a distance triangle can be used to work this out.



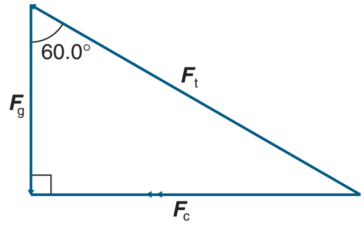
$r = 1.50 \times \sin 60.0°$

$r = 1.30$ m

b There are two forces acting—the tension in the cord, $\boldsymbol{F}_t$, and gravity, $\boldsymbol{F}_g$. The vertical upwards component of the tension is always equal to the downwards force due to gravity on the ball. The horizontal component of the tension force is unbalanced.



c The ball has an acceleration that is towards the centre of its circular path due to the unbalanced horizontal component of the tension force. This centripetal acceleration is horizontal and towards the centre of rotation. The centripetal force will also lie in this direction at this instant. A force triangle and trigonometry can be used here.



$m = 57.0 \times 10^{-3}$ g

$\boldsymbol{g} = 9.80$ m s^{-2}

$F_g = m\boldsymbol{g} = (57.0 \times 10^{-3})(9.80)$

$= 0.559$ N

$F_c = (0.559)\tan 60.0°$

$= 0.968$ N towards the centre

d Trigonometry can also be used to determine the tension in the cord.

$F_g = 0.559$ N

$\cos 60.0° = \dfrac{F_g}{F_t}$

$F_t = \dfrac{0.559}{\cos 60.0°}$

$= 1.12$ N along the cord

e $F_c = 0.968$ N

$m = 57.0 \times 10^{-3}$ kg

$r = 1.30$ m

$F_c = \dfrac{mv^2}{r}$

$v = \sqrt{\dfrac{F_c r}{m}} = \sqrt{\dfrac{(0.968)(1.30)}{57.0 \times 10^{-3}}}$

$= 4.70$ m s^{-1}

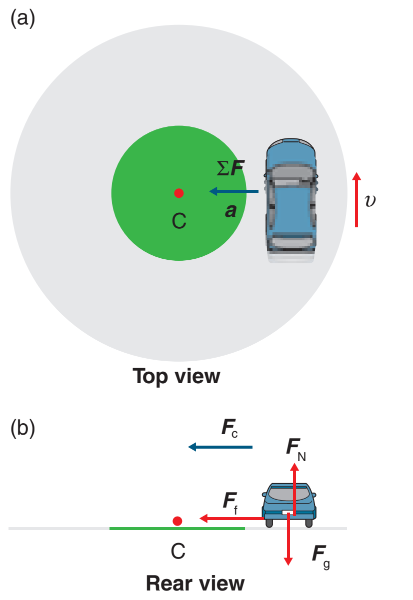


Figure 1.16

(a) The car is travelling in a circular path on a horizontal track. The centripetal acceleration and centripetal force are towards C. (b) The vertical forces balance each other and it is the frictional force between the tyres and the road that enables the car to corner.

Consider a car travelling anticlockwise around a horizontal roundabout with a constant speed v. As can be seen in Figure 1.16a, the car has a centripetal acceleration towards C and so centripetal force is also towards C. In Figure 1.16b, the forces acting on the car can be seen. The vertical forces (i.e. gravity and the normal force of the road on the car) are balanced. The only horizontal force is the sideways force that the road exerts on the car tyres. This is a force of friction, $\boldsymbol{F}_f$, and is unbalanced, so this is equal to the centripetal force, $\boldsymbol{F}_c$. If the car drove over an icy patch, there would be no friction and the car would not be able to turn. It would skid straight ahead in a line which is at a tangent to the circular path.

Banking the track eliminates the need for a sideways frictional force and allows the cars to travel faster without skidding out of the circular path. Consider the same car travelling around a circular, banked track with constant speed v. It is possible for the car to travel at a speed so that there is no sideways frictional force. This is called the design speed and it is dependent on the angle θ at which the track is banked. At this speed, the car exhibits no tendency to drift higher or lower on the track.

The car is still accelerating towards the centre of the circle C and so there must be an unbalanced force in this direction. Due to the banking, there are now only two forces acting on the car, its force due to gravity, $\boldsymbol{F}_g$, and the normal force, $\boldsymbol{F}_N$, from the track, as can be seen in Figure 1.17b. The normal force, $\boldsymbol{F}_N$ has a y-component which must oppose and balance the force due to gravity, as the car is not accelerating up or down, therefore $\boldsymbol{F}_{\text{Normal } y} = -\boldsymbol{F}_g$. The x-component of the normal force is unbalanced and provides the centripetal force that is horizontal and directed towards C, therefore $\boldsymbol{F}_{\text{Normal } x} = \boldsymbol{F}_c$. The x and y vector components add to give the normal force vector, $\boldsymbol{F}_N = \boldsymbol{F}_c + (-\boldsymbol{F}_g)$. If the angle and force due to gravity are known, trigonometry can be used to calculate the centripetal force (Figure 1.17c) and so determine the design speed. It is worth noting that the normal force will be larger here than on a flat track. The driver and car would *feel* heavier as there is a larger normal force acting on the driver and the car from the road.

The cyclist in Worked Example 1.2E, travelling on this section of track at 21.0 m s^{-1}, would ride perpendicular to the track and would experience no sideways forces up or down the track. It would be as though they were riding in a straight line on a flat track. Even if the track was made of ice, the cyclist could maintain their circular motion around the velodrome at this speed.

Interactive tutorial

High-speed cornering

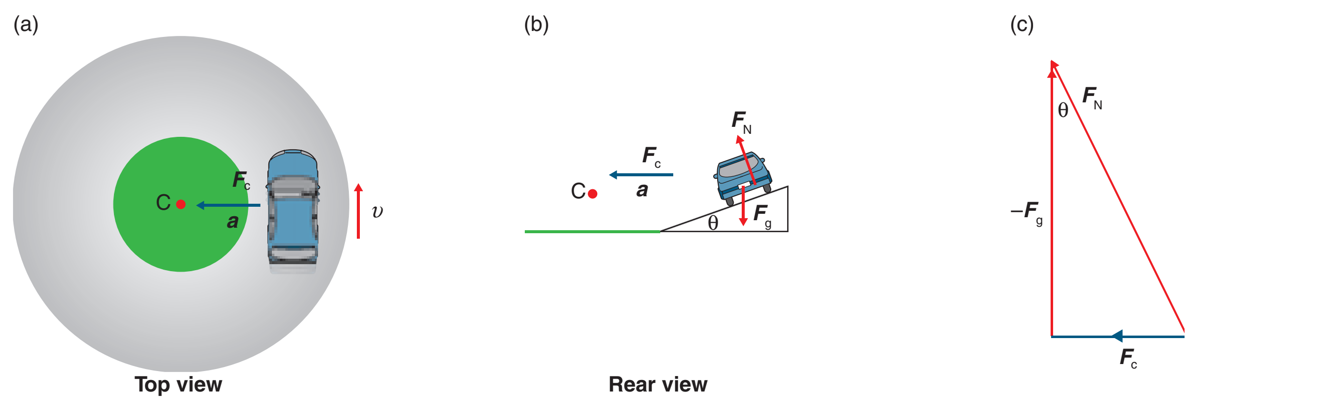


Figure 1.17

(a) The car is travelling in a circular path on a banked track. The centripetal acceleration and centripetal force are towards C. (b) The banked track means that the normal force has an inwards horizontal component. This is what enables the car to turn the corner. (c) Vector addition gives the centripetal force acting horizontally towards the centre.

However, what would happen if they slowed down? Gravity would not change, but now both the normal force and centripetal force would be smaller. The cyclist would have to depend on friction to stop them from moving down towards the bottom of the slope. They would need to lean the bike so that it was not perpendicular to the track, but was more vertically aligned.

Note that the mass of the cyclist is not a factor. If the mass was not known, the design speed could still be determined by leaving an unknown m in the calculations. This will cancel out.

✓ Worked Example 1.2E

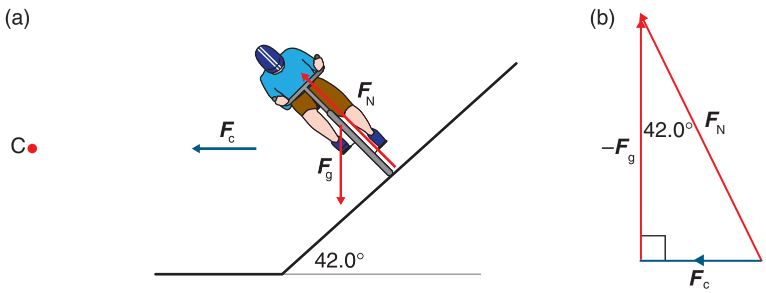
A curved section of track on an Olympic velodrome has radius of 50.0 m and is banked at an angle of 42.0° to the horizontal.

a Calculate the centripetal force acting on a cyclist riding at the design speed.

b At what speed would a cyclist of mass 75.0 kg need to travel if they were to experience zero frictional forces up or down the track, i.e. what is the design speed for this section of the velodrome?

Solution

a The forces acting on the cyclist are gravity and the normal force from the track. The centripetal force is the horizontal component of the normal force and is directed towards the centre of the circular track as shown in diagram (a) below.



Adding the force vectors gives a right-angle triangle as shown in diagram (b). Trigonometry can be used to work out the net force.

$m = 75.0$ kg $\qquad \tan\theta = \dfrac{F_c}{-F_g}$

$g = -9.80$ m s^{-1} $\qquad F_c = -mg\tan\theta = -(75.0)(-9.80)(\tan 42.0°)$

$= 6.62 \times 10^2$ N towards the centre of curvature

b The centripetal force is used to find the speed:

$m = 75.0$ kg $\qquad F_c = \dfrac{mv^2}{r}$

$r = 50.0$ m $\qquad v = \sqrt{\dfrac{(6.62 \times 10^2)(50.0)}{75.0}}$

$F_c = 6.62 \times 10^2$ N $\qquad = 21.0$ m s^{-1}

Physics file

Figure 1.18
For a rider to successfully conquer the Wall of Death, they would need to travel reasonably fast and there would need to be good grip between the tyres and the track. The rider is relying on friction to maintain their motion halfway up the wall.

In some amusement parks in other parts of the world, there is a ride known menacingly as the Wall of Death. It consists of a cylindrical enclosure with vertical walls. People on motorbikes ride into the enclosure and around the vertical walls, so the angle of banking is 90°! The riders need to keep moving and are depending on friction to hold them up. By travelling fast, the centripetal force (the normal force from the wall) is large and this increases the size of the grip (friction) between the wall and tyres. If the rider slammed on the brakes and stopped, they would simply plummet.

Physics in action — Leaning into corners

In many sporting events, the participants need to travel around corners at high speeds. Motorbike riders lean their bikes over almost to the track as they corner. This leaning technique is also evident in ice skating, bicycle races, skiing and even when you run round a corner. It enables the competitor to corner at high speed without falling over. Why is this so?

Consider a bike rider cornering on a horizontal road surface. The forces acting on the bike and rider (Figure 1.19b) are unbalanced. The forces are the force due to gravity, $\boldsymbol{F}_g$, and the force from the track. The track exerts a reaction force, $\boldsymbol{F}_r$, on the rider that acts both inwards and upwards. The inwards component is the frictional force, $\boldsymbol{F}_f$, between the track and the tyres. The upwards component is the opposite of the weight force, $-\boldsymbol{F}_g$, from the track.

The rider is travelling in a horizontal circular path at constant speed, and so has a centripetal acceleration directed towards the centre of the circle at C. Therefore, the centripetal force is directed towards C. By analysing the vertical and horizontal components in Figure 1.19b, we see that the force due to gravity, $\boldsymbol{F}_g$, must balance the opposite of the weight force, $-\boldsymbol{F}_g$. The centripetal force that is producing the centripetal acceleration is supplied by the frictional force, $\boldsymbol{F}_f$. In other words, the rider is depending on a sideways frictional force to turn the corner. An icy or oily patch on the track would cause the tyres to slide out from under the rider, and he or she would slide painfully along the road at a tangent to the circular path.

(a)

(b)

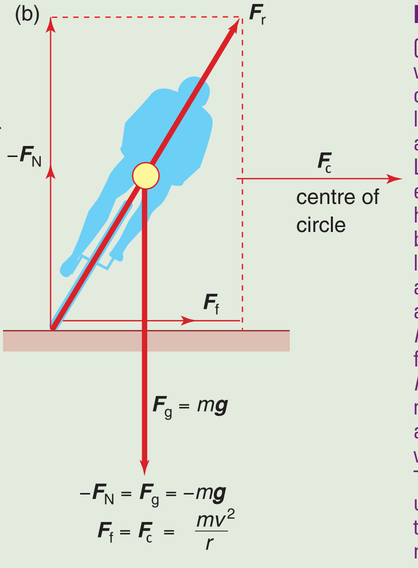


Figure 1.19 (a) Australia's Casey Stoner won the 2007 MotoGP championship. Here he is leaning his bike as he takes a corner at Phillip Island. Leaning into the corner enables him to corner at higher speeds. In fact, the bike would flip if he did not lean it. (b) The forces acting as the rider turns a corner are the force due to gravity, F_g, the opposite of the weight force, $-F_g$, and the friction, F_f, between the tyres and the road. The force due to gravity and the opposite of the weight force are balanced. The friction supplies the unbalanced centripetal force that leads to the circular motion.

1.2 SUMMARY Circular motion in a horizontal plane

- An object moving with a uniform speed in a circular path of radius r and with a period T has an average speed that is given by $v = \frac{2\pi r}{T}$.
- The velocity of an object moving with a constant speed in a circular path is continually changing and is at a tangent to the circular path.
- An object moving in a circular path with a constant speed has an acceleration due to its circular motion. This acceleration is directed towards the centre of the circular path and is called centripetal acceleration:

$$\boldsymbol{a}_c = \frac{v^2}{r} = \frac{4\pi^2 r}{T^2}$$

- Centripetal acceleration is a consequence of a centripetal force acting to make an object move in a circular path.
- Centripetal forces are directed towards the centre of the circle and their magnitude can be calculated by using Newton's second law:

$$\boldsymbol{F}_c = m\boldsymbol{a}_c = \frac{mv^2}{r} = \frac{m4\pi^2 r}{T^2}$$

- Centripetal force is always supplied by a real force, the nature of which depends on the situation. The real force is commonly friction, gravitation or the tension in a string or cable.

1.2 SUMMARY Circular motion in a horizontal plane cont.

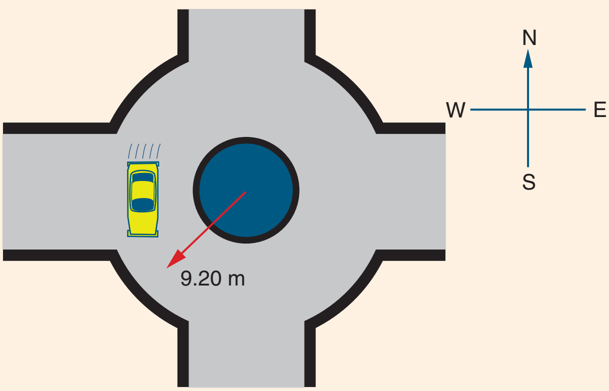
- An object can move in a circular path attached to string or cable that is not horizontal. This is called a conical pendulum. Here the centre of the circular path is not the end of the string, but in the same horizontal plane as the object itself.
- In a conical pendulum, the centripetal force is the horizontal component of the tension. The vertical component of the tension balances the force due to gravity on the object. The acceleration of the object is horizontal and towards the centre of the circular path.
- The banking of a track is where the track is inclined at some angle to the horizontal. This enables vehicles to travel at higher speeds as they corner.
- Banking a track eliminates the need for a sideways frictional force to turn. When the speed and angle are such that there is no sideways frictional force, the speed is known as the design speed.
- The forces acting on a vehicle travelling at the design speed on a banked track are gravity and the normal force from the track. The vertical component of the normal force balances the force due to gravity, while the horizontal component gives the centripetal force towards the centre of the circular motion.

1.2 Questions

In the following questions, assume that the acceleration due to gravity is 9.80 m s^{-2} and ignore the effects of air resistance.

The following information applies to questions 1–5.

A car of mass 1.20×10^3 kg is travelling on a roundabout in a circular path of radius 9.20 m. The car moves with a constant speed of 8.00 m s^{-1}.



1 a Which two of the following statements correctly describe the motion of the car as it travels around the roundabout?
   A It has a constant speed.
   B It has a constant velocity.
   C It has zero acceleration.
   D It has an acceleration that is directed towards the centre of the roundabout.

  b As the car turns towards the left, a passenger describes the effect on her as 'being thrown across towards the right side of the cabin'. What has actually happened?

2 When the car is in the position shown in the diagram, what is the:
  a speed of the car?
  b velocity of the car?
  c magnitude and direction of the acceleration of the car?

3 a Calculate the magnitude and direction of the net force acting on the car at the position shown.
  b Identify the force that is enabling the car to move in its circular path.

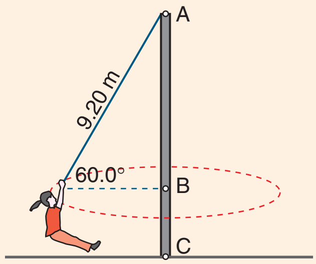
4 Some time later, the car has travelled halfway around the roundabout. What is:
  a the velocity of the car at this point?
  b the direction of its acceleration at this point?

5 If the driver of the car kept speeding up, what would eventually happen to the car as it travelled around the roundabout? Explain.

6 An ice skater of mass 50.0 kg is skating in a horizontal circle of radius 1.50 m at a constant speed of 2.00 m s^{-1}.
  a What is the acceleration of the skater?
  b Are the forces acting on the skater balanced or unbalanced? Explain.
  c Calculate the magnitude of the net force acting on the skater.
  d Identify the force that is enabling the skater to move in a circular path.

7 An athlete competing at a junior sports meet swings a 2.50 kg hammer in a horizontal circle of radius 0.800 m at 2.00 revolutions per second. Assume that the wire is horizontal at all times.
  a What is the period of rotation of the ball?
  b What is the orbital speed of the ball?
  c Calculate the acceleration of the ball.
  d What is the magnitude of the net force acting on the ball?
  e Name the force that is responsible for the centripetal acceleration of the ball.
  f Describe the motion of the ball if the wire breaks.

8 Ella of mass 30.0 kg is playing on a maypole swing in a playground. The rope is 2.40 m long and at an angle of 60.0° to the horizontal as she swings freely in a circular path. Ignore the mass of the rope in your calculations.



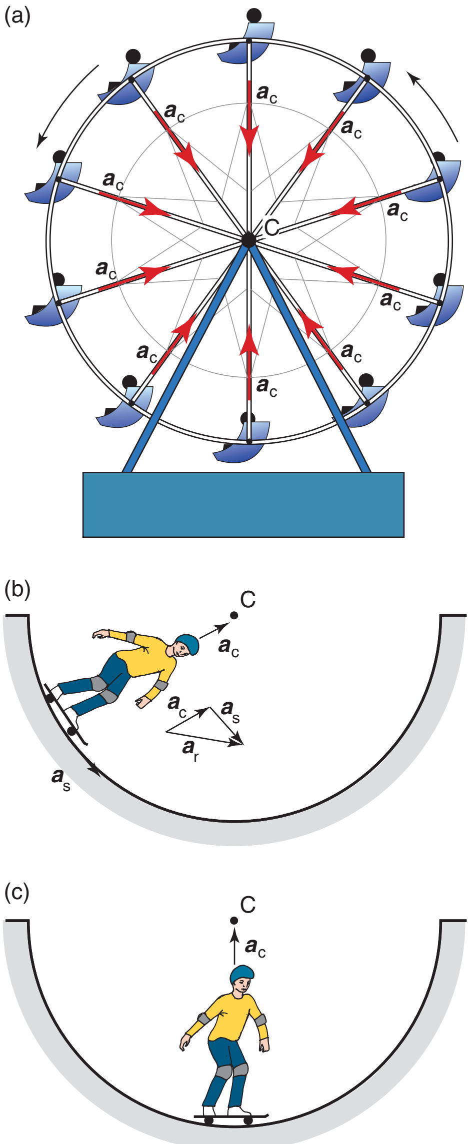
a Calculate the radius of her circular path.
b Identify the forces that are acting on her as she swings freely.
c What is the direction of her acceleration when she is at the position shown in the diagram?
d Calculate the net force acting on the girl.
e What is her speed as she swings?

The following information applies to questions 9 and 10.

A section of track at a NASCAR raceway is banked to the horizontal. The track section is circular with a radius of 80 m and design speed 18 m s^{-1}. A car of mass 1.20×10^3 kg is being driven around the track at 18.0 m s^{-1}.

9 a Calculate the magnitude of the net force acting on the car (in kN).
b Calculate the angle to the horizontal at which the track is banked.

10 The driver now drives around the track at 30.0 m s^{-1}. What would the driver have to do maintain their circular path around the track?



1.3 Circular motion in a vertical plane

When you go on a roller-coaster ride you will travel over humps and down dips at high speeds. There are also fairground rides that take you through full vertical circles. During these rides, your body will experience forces that you may or may not find pleasant.

We saw in section 1.2 that a body moving with constant speed in a circular path has an acceleration that is directed towards the centre of the circle. The same applies even if, as on a ferris wheel, the body is moving in a vertical circle (Figure 1.20a).

Circular motion in a vertical plane, however, is not usually uniform; this is because the speed of the body varies. An example of this is a skateboarder practising in a half-pipe. The speed of the skateboarder will increase on the way down as there is a component of the acceleration due to gravity acting down the slope. This means there will be a component of gravitational acceleration down the slope, $\boldsymbol{a}_s$, as well as a centripetal acceleration, $\boldsymbol{a}_c$. The resultant acceleration $\boldsymbol{a}_r$ is the vector addition of these two accelerations (Figure 1.20b). At the bottom of the 'pipe', however, the skateboarder will be neither slowing down nor speeding up, so the acceleration is purely centripetal at this point (Figure 1.20c). The same applies at the very top of a circular path. For this reason, motion at these points is easier to analyse.

Figure 1.20
(a) The seats in a ferris wheel move with a constant speed and so have a constant centripetal acceleration directed towards the centre C. (b) The skateboarder is speeding up and so has both a linear and a centripetal acceleration. The resultant acceleration, $\boldsymbol{a}_r$, is not towards C. (c) At the lowest point the speed of the skateboarder is momentarily constant, so there is no linear acceleration. The acceleration is supplied completely by the centripetal acceleration, and is towards C.

Theme-park rides make you appreciate that the forces you experience throughout the ride can vary greatly. To illustrate this, consider the case of a person in a roller-coaster cart travelling horizontally at 4.00 m s^{-1}. If the person's mass is 50.0 kg and the gravitational field strength is –9.80 m s^{-2}, the forces acting on the person can be calculated. These forces are the force due to gravity, $\boldsymbol{F}_g$, and the normal reaction force provided by the seat, $\boldsymbol{F}_{\text{seat on person}}$ (Figure 1.21). The person is moving in a straight line with a constant speed and so there is no unbalanced force acting. The force due to gravity is equal and opposite to the normal reaction force from the seat so they balance. Using the convention that up is positive and down is negative the normal force provided by the seat is therefore $+4.90 \times 10^2$ N, while the force due to gravity is -4.90×10^2 N. The most important force involved in roller-coaster rides is the reaction force of the seat on the rider, as that is the force that determines what the rider will feel. It is the force that makes you feel like you are stuck to the seat as well as providing that thrilling feeling that you and your seat are about to part company!

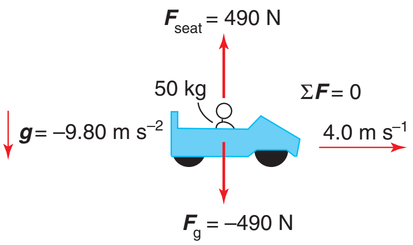


Figure 1.21
The vertical forces are balanced in this situation, i.e. $\boldsymbol{F}_{seat} = -\boldsymbol{F}_g$.

Now consider the force of the seat that acts on the person after the cart has reached the bottom of a circular dip of radius 2.50 m and is now moving at 8.00 m s^{-1} (Figure 1.22). The person will have a centripetal acceleration due to the circular path that they must follow. This centripetal acceleration is due to a centripetal force directed towards the centre C of the circular path—in this case, vertically upwards, in the positive direction. The person's centripetal force is:

$$\updownarrow^{+}_{-} \qquad \boldsymbol{F}_c = \frac{mv^2}{r}$$

$$m = 50.0 \text{ kg} \qquad \boldsymbol{F}_c = \frac{(50.0)(8.00)^2}{2.50}$$

$$v = 8.00 \text{ m s}^{-1} \qquad = +1.28 \times 10^3 \text{ N}$$

$$r = 2.50 \text{ m} \qquad = 1.28 \times 10^3 \text{ N upwards}$$

At the bottom of the curve, not only does the seat need to provide the upwards force to balance the force due to gravity, but it must also provide the centripetal force to make the rider travel a circular path. The reaction force of the seat on the rider, $\boldsymbol{F}_{seat}$, is the vector sum of the centripetal force $\boldsymbol{F}_c$, and the opposite to the force due to gravity $-\boldsymbol{F}_g$.

$$\updownarrow^{+}_{-} \qquad \boldsymbol{F}_{seat} = \boldsymbol{F}_c + (-\boldsymbol{F}_g)$$

$$\boldsymbol{F}_c = +1.28 \times 10^3 \text{ N} \qquad = (+1.28 \times 10^3) + \big(-(50.0)(-9.80)\big)$$

$$g = -9.80 \text{ m s}^{-2} \qquad = +1.77 \times 10^3 \text{ N}$$

$$m = 50.0 \text{ kg} \qquad = 1.77 \times 10^3 \text{ N upwards}$$

They add together to give an upwards force of 1.77×10^3 N. This indicates that the reaction force of the seat on the rider must be greater than the rider's weight force, in other words, the rider feels heavier than they normally do and they feel like they are pressing into their seat firmly. In actual fact it is the seat that is pressing into the rider quite firmly, in order to change the momentum of the rider into a curved path.

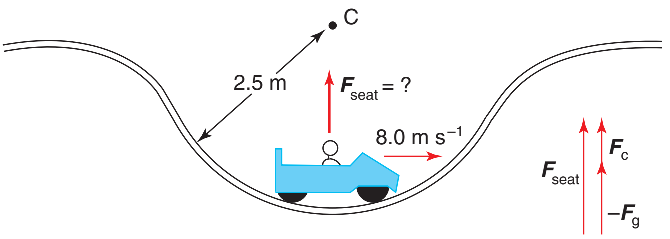


Figure 1.22
The person has a centripetal acceleration that is directed upwards, and so the centripetal force is also upwards. So here, the magnitude of the normal force, $\boldsymbol{F}_{\text{seat}}$, is greater than the opposite of the force due to gravity, $-\boldsymbol{F}_g$, as it must also provide the centripetal force. This is a situation that is felt by the rider as feeling heavier than normal.

Now consider the situation as the cart moves over the top of a hump of radius 2.50 m with a lower speed of 2.00 m s^{-1} (Figure 1.23). The person now has a centripetal acceleration that is directed vertically downwards towards the centre C of the circle. Therefore, the centripetal force acting at this point is also directed vertically downwards. The reaction force of the seat on the rider is:

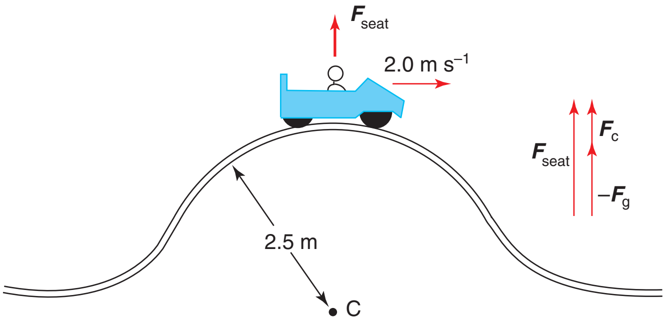


Figure 1.23
The centripetal acceleration is downwards, and so the centripetal force is also in that direction. At this point, the magnitude of the reaction force of the seat on the person, $\boldsymbol{F}_{\text{seat}}$, is less than the weight, $\boldsymbol{F}_g$, of the person.

$\updownarrow^{+}_{-}$	$\boldsymbol{F}_{\text{seat}} = \boldsymbol{F}_c + (-\boldsymbol{F}_g)$
$v = 2.00 \text{ m s}^{-1}$	$= \dfrac{mv^2}{r} + (-mg)$
$r = -2.50 \text{ m}$	$= \left(\dfrac{(50.0)(2.00)^2}{-2.50}\right) + \left(-(50.0)(-9.80)\right)$
$g = -9.80 \text{ m s}^{-2}$	$= (-8.00 \times 10^1) + (+4.90 \times 10^2)$
$m = 50.0 \text{ kg}$	$= +4.10 \times 10^2 \text{ N}$
	$= 4.10 \times 10^2 \text{ N upwards}$

The force due to gravity never changes as the mass of the rider never changes, therefore the opposite of the force due to gravity is also constant. The centripetal force is now downwards, the negative radius enables the centripetal force vector to become negative too. Here the reaction force of the seat on the rider is less than is would normally be, so the rider will feel like they are not pushing as hard on their seat. The rider will feel this as being lighter than they normally feel.

Physics file

A fighter pilot in a loop manoeuvre can safely experience centripetal accelerations of up to around 5*g*, or 49 m s^{-2}. In a loop where the g forces are greater than this, the pilot may pass out. If the pilot flies with his or her head inside the loop, the centripetal acceleration of the plane will make the blood flow away from the head. The resulting lack of blood in the brain may cause the pilot to lose consciousness ('black out'). Fighter pilots wear 'g suits', which pressurise the legs to prevent blood flowing into them.
On the other hand, if the pilot's head is on the outside of the loop, the additional blood flow to the head can make the whites of the eyes turn red. The excess blood flow in the head may cause 'red out'.

The weight of the person has not changed, but they feel heavier and lighter as they travel through the dips and humps. This is because the reaction force of the seat acting on them varies throughout the ride. The reaction force acting on the person gives their *apparent weight*. This will be discussed in the next chapter.

Physics in action — Lift-off!

During a car rally it is quite common to see a car travelling at high speed become airborne as it travels over a rise in the road. This occurs because the speed of the car is too great for the radius of curvature of the road. The same thing would happen to the person in the roller-coaster cart we looked at previously if its speed was great enough and the person wasn't strapped in. Imagine the roller-coaster cart is now travelling over the same rise as in Figure 1.23, but at an increased speed: 4.95 m s⁻¹. This speed would give a reaction force of:

$\updownarrow^{+}_{-}$ $\quad \boldsymbol{F}_{seat} = \boldsymbol{F}_g + (-\boldsymbol{F}_g)$

$v = 4.95 \text{ m s}^{-1}$ $\quad \boldsymbol{F}_{seat} = \frac{mv^2}{r} + (-mg)$

$r = -2.50 \text{ m}$ $\quad \boldsymbol{F}_{seat} = \left(\frac{(50.0)(4.95)^2}{(-2.50)}\right) + \left(-(50.0)(-9.80)\right)$

$g = -9.80 \text{ m s}^{-2}$ $\quad \boldsymbol{F}_{seat} = (-4.90 \times 10^2) + (+4.90 \times 10^2)$

$m = 50.0 \text{ kg}$ $\quad = 0 \text{ N}$

By considering the force vectors, the force due to gravity fully accounts for the centripetal force. In other words, the reaction force of the seat on the rider, $\boldsymbol{F}_{seat}$, is now zero. While travelling over the hump, there would be no interaction between the person and the seat, and they would be at the point of lifting off. If the speed of the cart had been any greater, the person *would* have left their seat and would have needed a restraint to remain in the cart.

Figure 1.24
When a car travels too fast for the radius of curvature of the road, the reaction force of the road on the car, $\boldsymbol{F}_{road}$, is reduced to zero and the car will just lift off the road surface.

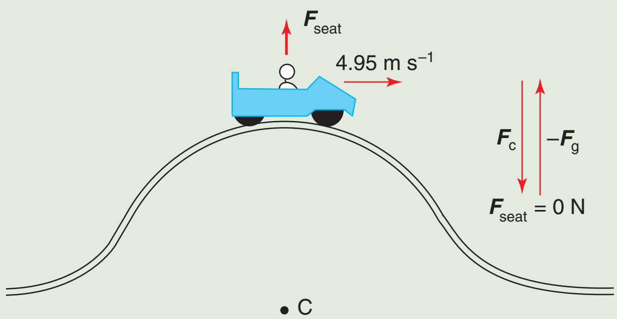
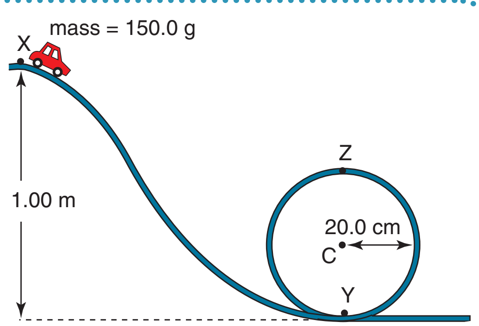


Figure 1.25
At this speed the centripetal force, $\boldsymbol{F}_c$, is equal to the weight, $\boldsymbol{F}_g$, of the person. In other words, the normal reaction force of the seat on the person, $\boldsymbol{F}_{seat}$, is zero. The person would feel as though they were at the point of lifting off their seat.

✓ Worked Example 1.3A

A student arranges a toy car track with a vertical loop of radius 0.200 m, as shown.
A car of mass 150.0 g is released from a height of 1.00 m at point X. The car rolls down the track and travels around the loop. Assuming $\boldsymbol{g}$ is −9.80 m s⁻², and ignoring friction, calculate the:

a speed of the car as it reaches the bottom of the loop, point Y
b centripetal acceleration of the car at point Y
c reaction force of the track on the car at point Y
d speed of the car as it reaches the top of the loop at point Z
e reaction force of the track on the car at point Z
f release height from which the car will just maintain contact with the track as it travels past point Z.



Solution

a At point X, the car's total energy is given by:

$m = 150.0 \times 10^{-3}$ kg

$\boldsymbol{g} = 9.80$ m s^{-2}

$v = 0$ m s^{-1}

$\Delta \boldsymbol{h} = 1.00$ m

$$E_{total} = E_k + E_p$$
$$= \tfrac{1}{2}mv^2 + m\boldsymbol{g}\Delta\boldsymbol{h}$$
$$= (0) + (150.0 \times 10^{-3})(9.80)(1.00)$$
$$= 1.47 \text{ J}$$

Note that this problem ignores the sign convention when calculating the scalar quantity – energy. As the car rolls down the track, it loses its gravitational potential energy and gains kinetic energy. At the bottom of the loop, the car has zero potential energy. Therefore its kinetic energy at Y is 1.47 J. Its speed can now be determined.

$m = 150.0 \times 10^{-3}$ kg

$E_k = 1.47$ J

$$E_k = \tfrac{1}{2}mv^2$$
$$v = \sqrt{\frac{2E_k}{m}} = \sqrt{\frac{2(1.47)}{150.0 \times 10^{-3}}}$$
$$= 4.43 \text{ m s}^{-1}$$

b The centripetal acceleration of the toy car at Y is:

$\updownarrow^{+}_{-}$

$r = +20.0 \times 10^{-2}$ m

$v = 4.43$ m s^{-1}

$$\boldsymbol{a}_c = \frac{v^2}{r}$$
$$= \frac{(4.43)^2}{20.0 \times 10^{-2}}$$
$$= +98.0 \text{ m s}^{-2}$$
$$= 98.0 \text{ m s}^{-2} \text{ upwards}$$

c The reaction force of the track on the car at Y is:

$\updownarrow^{+}_{-}$

$\boldsymbol{a}_c = +98.0$ m s^{-2}

$m = 150.0 \times 10^{-3}$ kg

$g = -9.80$ m s^{-2}

$$\boldsymbol{F}_{track} = \boldsymbol{F}_c + (-\boldsymbol{F}_g)$$
$$= m\boldsymbol{a}_c + (-mg)$$
$$= (150.0 \times 10^{-3})(98.0) + \left(-(150.0 \times 10^{-3})(-9.80)\right)$$
$$= 14.7 + (+1.47)$$
$$= 16.2 \text{ N upwards}$$

This is the apparent weight of the car, which at point Y is over 10 times greater than its actual weight.

d As the car travels up to point Z, it loses kinetic energy and gains gravitational potential energy. Its total energy, however, remains 1.47 J. Point Z is at a height of 0.40 m, so the remaining kinetic energy is:

$m = 150.0 \times 10^{-3}$ kg

$\boldsymbol{g} = 9.80$ m s^{-2}

$v = 0$ m s^{-1}

$\Delta \boldsymbol{h} = 0.400$ m

$E_{total} = 1.47$ J

$$E_{k\text{ at Z}} = E_{total} - E_{p\text{ at Z}}$$
$$\tfrac{1}{2}mv^2 = E_{total} - (m\boldsymbol{g}\Delta\boldsymbol{h})$$
$$v = \sqrt{\frac{2\left(E_{total} - (m\boldsymbol{g}\Delta\boldsymbol{h})\right)}{m}}$$
$$= \sqrt{\frac{2\left(1.47 - (150.0 \times 10^{-3})(9.80)(0.400)\right)}{150.0 \times 10^{-3}}}$$
$$= 3.43 \text{ m s}^{-1}$$

e The reaction force of the track on the car at Z is:

$\updownarrow^{+}_{-}$

$$\boldsymbol{F}_{\text{track on car}} = \boldsymbol{F}_c + (-\boldsymbol{F}_g)$$

$v = 3.43 \text{ m s}^{-1}$ $\qquad = \frac{mv^2}{r} + (-mg)$

$r = -20.0 \times 10^{-2} \text{ m}$ $\qquad = \left(\frac{(150.0 \times 10^{-3})(3.43)^2}{-20.0 \times 10^{-2}}\right) + \left(-(150.0 \times 10^{-3})(-9.80)\right)$

$g = -9.80 \text{ m s}^{-2}$ $\qquad = (-8.82 \times 10^2) + (+1.47 \times 10^2)$

$m = 150.0 \times 10^{-3} \text{ kg}$ $\qquad = -7.35 \text{ N}$

$\qquad = 7.35 \text{ N downwards}$

This is about five times greater than the actual weight.

f When the car is travelling at the speed at which it just loses contact with the track, the reaction force of the track on the car, $\boldsymbol{F}_{\text{track on car}}$, is zero.

$\updownarrow^{+}_{-}$

$$\boldsymbol{F}_{\text{track on car}} = \boldsymbol{F}_c + (-\boldsymbol{F}_g)$$

$\boldsymbol{F}_{\text{track on car}} = 0 \text{ N}$ $\qquad \frac{mv^2}{r} = mg$

$r = -20.0 \times 10^{-2} \text{m}$ $\qquad v = \sqrt{rg}$

$g = -9.80 \text{ m s}^{-2}$ $\qquad = \sqrt{(-20.0 \times 10^{-2})(-9.80)}$

$\qquad = 1.40 \text{ m s}^{-1}$

The total energy of the toy car at point Z is therefore:

$m = 150.0 \times 10^{-3} \text{ kg}$ $\qquad E_{\text{total}} = E_k + E_p$

$\boldsymbol{g} = 9.80 \text{ m s}^{-2}$ $\qquad = \frac{1}{2}mv^2 + m\boldsymbol{g}\Delta\boldsymbol{h}$

$v = 1.40 \text{ m s}^{-1}$ $\qquad = \frac{1}{2}(150.0 \times 10)(1.40)^2 + (150.0 \times 10^{-3})(9.80)(0.400)$

$\Delta\boldsymbol{h} = 0.400 \text{ m}$ $\qquad = 0.735 \text{ J}$

So when the car is released, its height needs to be such that it has 0.735 J of gravitational potential energy:

$m = 150.0 \times 10^{-3} \text{ kg}$ $\qquad E_p = m\boldsymbol{g}\Delta\boldsymbol{h}$

$\boldsymbol{g} = 9.80 \text{ m s}^{-2}$ $\qquad \Delta\boldsymbol{h} = \frac{E_p}{m\boldsymbol{g}} = \frac{0.735}{(150.0 \times 10^{-3})(9.80)}$

$E_p = 0.735 \text{ J}$ $\qquad = 0.500 \text{ m}$

The student should release the car from a height of 50.0 cm for it to complete the loop.

Physics in action — How to travel upside down without falling out

You might have been on a roller-coaster like the one in Figure 1.26, where you were actually upside down at times during the ride. These rides use their speed and the radius of their circular path to prevent the riders from falling out. In theory, the safety harnesses worn by the riders are not needed to hold the people in their seats.

Figure 1.26

The thrill seekers on this roller-coaster ride do not fall out when upside down because the centripetal acceleration of the cart is greater than 9.8 m s^{-2} down.

The reason people do not fall out is that their centripetal acceleration while on the roller-coaster is greater than the acceleration due to gravity (−9.80 m s^{-2}). To understand the significance of this, try the following activity. Place an eraser on the palm of your hand, then turn your hand palm down and move it rapidly towards the floor. You should find, after one or two attempts, that it is possible to keep the eraser in contact with your hand as you 'push' it down. The eraser is upside down, but it is not falling out of your hand. Your hand must be moving down with an acceleration in excess of −9.80 m s^{-2}, and continually exerting a normal force on the eraser. This acceleration of 9.80 m s^{-2} down is the critical point in this exercise. If your hand had an acceleration less than this, the eraser would fall away from your hand to the floor. A similar principle holds with roller-coaster rides. The people on the ride do not fall out at the top because the motion of the roller-coaster gives them a centripetal acceleration that is greater than 9.80 m s^{-2} down. The engineers who designed the ride would have ensured that the roller-coaster moves with sufficient speed and in a circle of the appropriate radius so that this happens.

As an example, consider a ride of radius 15 m in a simple vertical circle (Figure 1.27). It is possible to calculate the speed that would ensure that a rider cannot fall out. We will assume that the person has a mass of 45 kg and that g is −9.80 m s^{-2}. At the critical speed, the reaction force of the seat, $\boldsymbol{F}_{seat}$, on the person will be zero. In other words, the seat will exert no force on the person at this speed:

$$\updownarrow^{+}_{-} \qquad \boldsymbol{F}_{\text{seat on person}} = \boldsymbol{F}_c + (-\boldsymbol{F}_g)$$

$$\boldsymbol{F}_{\text{seat on person}} = 0\text{ N} \qquad \frac{mv^2}{r} = mg$$

$$r = -15.0\text{ m} \qquad v = \sqrt{rg}$$

$$g = -9.80\text{ m s}^{-2} \qquad = \sqrt{(-15.0)(-9.80)}$$

$$= 12.1\text{ m s}^{-1}$$

This speed is equal to 43 km h^{-1} and is the minimum needed to prevent the riders from falling out. In practice, the roller-coaster would move with a speed much greater than this to ensure that there was a significant force between the patrons and their seats. Corkscrew roller-coasters can travel at up to 110 km h^{-1} and the riders can experience accelerations of up to 50 m s^{-2} (5g). So, safety harnesses are really only needed when the speed is below the critical value; their primary function is to prevent people from moving around while on the ride.

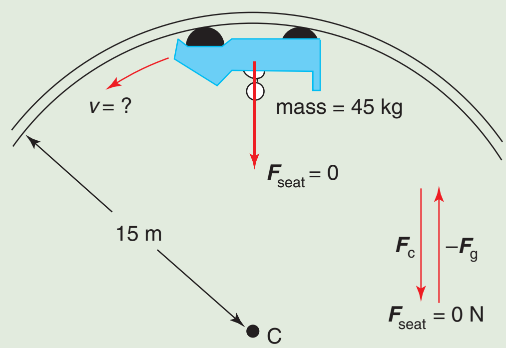


Figure 1.27

At the speed at which the reaction force of the seat on the person, F_{seat}, becomes zero, the centripetal force, F_c, on the rider will equal the force due to gravity, F_g, and they will move so that the centripetal acceleration is 9.80 m s^{-2} down.

1.3 SUMMARY Circular motion in a vertical plane

- Objects that are moving in vertical circular paths experience a centripetal acceleration and centripetal force that acts towards the centre of their path.
- The gravitational force must be considered when analysing the motion of an object moving in a vertical circle.
- At the point where a moving object lifts off from its circular path, the object will be moving with a centripetal acceleration that is equal to that due to gravity.
- The apparent weight of a body is given by the normal reaction force that is acting on the body from a surface, like a seat or track. This may be different from the actual weight.

1.3 Questions

Assume that $g = -9.80$ m s^{-2} and ignore the effects of air resistance.

The following information applies to questions 1 and 2.

A yo-yo is swung with a constant speed in a vertical circle.

1 a Describe the acceleration of the yo-yo in its path.
 b At which point in the circular path is there the greatest amount of tension in the string?
 c At which point in the circular path is there the lowest amount of tension in the string?
 d At which point is the string most likely to break?

2 If the yo-yo has a mass of 80.0 g and the radius of the circle is 1.50 m, find the minimum speed that this yo-yo must have at the top of the circle so that the cord does not slacken.

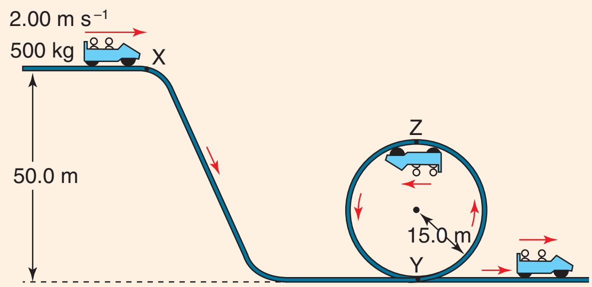
3 A car of mass 800.0 kg encounters a speed hump of radius 10.0 m. The car drives over the hump at a constant speed of 14.4 km h^{-1}.
 a Name all the forces acting on the car when it is at the top of the hump.
 b Calculate the resultant force acting on this car when it is at the top of the hump.
 c After travelling over the hump, the driver remarked to a passenger that she felt lighter as the car moved over the top of the speed hump. Is this possible? Explain your answer.
 d What is the maximum speed (in km h^{-1}) that this car can have at the top of the hump and still have its wheels in contact with the road?

The following information applies to questions 4 and 5.

A popular amusement park ride is the 'loop-the-loop' in which a cart descends a steep incline at point X, enters a circular rail track at point Y, and makes one complete revolution of the circular track. The car, whose total mass is 500.0 kg, carries the passengers with a speed of 2.00 m s^{-1} when it begins its descent at point X from a vertical height of 50.0 m.

4 a Calculate the speed of the car at point Y.
 b What is the speed of the car at point Z?
 c Calculate the normal force acting on the car at Z.

5 What is the minimum speed that the car can have at point Z and still stay in contact with the rails?



The following information applies to questions 6 and 7.

A stunt pilot appearing at an air show decides to perform a vertical loop so that she is upside down at the top of the loop. During the stunt she maintains a constant speed of 35 m s^{-1} while completing the 100 m radius loop.

6 Calculate the apparent weight of the 80.0 kg pilot when she is at the top of the loop.

7 What minimum speed would the pilot need at the top of the vertical loop in order to experience zero normal force from the seat (i.e. to feel weightless)?

8 The maximum value of acceleration that the human body can safely tolerate for short time intervals is nine times that due to gravity. Calculate the maximum speed with which a pilot could safely pull out of a circular dive of radius 400.0 m.

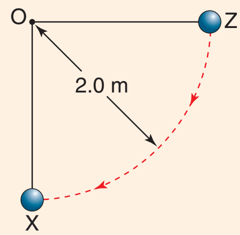
The following information applies to questions 9 and 10.

A light wire of length 2.00 m supports a bowling ball of mass 4.00 kg in a vertical position at point X as shown in the diagram.

9 **a** Calculate the tension in the wire when the ball is at rest at point X.

b The ball is then moved to point Z and released. Calculate the tension in the wire as the ball now moves through point X.

10 When is the wire more likely to break: when the ball is stationary at X, or when it is moving through X? Explain your reasoning.



Chapter 1 Review

For the following questions, assume that the acceleration due to gravity is 9.80 m s^{-2} and ignore the effects of air resistance unless otherwise stated. Questions indicated by * relate to material on the CD.

The following information applies to questions 1 and 2.

Hannah is jumping on her trampoline. Her mass is 55.0 kg and she lands vertically at 3.50 m s^{-1} before rebounding vertically at 3.00 m s^{-1}.

***1** **a** What is Hannah's change of speed as she bounces?

b What is Hannah's change of velocity as she bounces?

***2** Each time Hannah bounces, the trampoline exerts a force on her.

a Discuss this force and how it varies during the short duration of each bounce. What name is usually given to this force?

b Each time Hannah bounces, she exerts a downwards force on the trampoline. Which one of the following statements is correct?

A The force that Hannah exerts on the trampoline is always greater than the force that the trampoline exerts on Hannah.

B The force that Hannah exerts on the trampoline is always less than the force that the trampoline exerts on Hannah.

C The force that Hannah exerts on the trampoline is at times greater than and at other times less than the force that the trampoline exerts on Hannah.

D The force that Hannah exerts on the trampoline is always equal to the force that the trampoline exerts on Hannah.

***3** A motorcyclist is travelling at 60.0 km h^{-1} when she has to brake to a sudden stop. She skids and stops in a distance of 15.0 m. The combined mass of the bike and rider is 120.0 kg.

a What was the magnitude of her average acceleration as she stopped?

b How fast was she travelling after skidding for 1.50 s?

c Determine the magnitude of the average retarding force (in kN) acting on her motorbike as it stopped.

4 Two identical tennis balls X and Y are hit horizontally from a point 2.00 m above the ground with different initial speeds: ball X has an initial speed of 5.00 m s^{-1} while ball Y has an initial speed of 7.50 m s^{-1}.

a Calculate the time it takes for each ball to strike the ground.

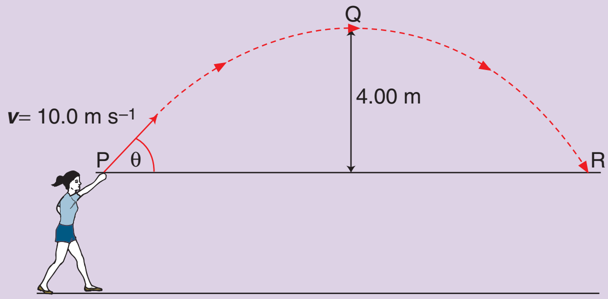
b Calculate the speed of ball X just before it strikes the ground.

c What is the speed of ball Y just before it strikes the ground?

d How much further than ball X does ball Y travel in the horizontal direction before bouncing?

The following information applies to questions 5–8.

The diagram shows the trajectory of a Vortex after it has been thrown with an initial speed of 10.0 m s^{-1}. The Vortex reaches its maximum height at point Q, 4.00 m higher than its starting height.



5 What is the value of the angle – that the initial velocity vector makes with the horizontal?

6 What is the speed of the Vortex at point Q?

7 What is the acceleration of the Vortex at point Q?

A zero

B 9.80 m s^{-2} forwards

C 4.90 m s^{-2} down

D 9.80 m s^{-2} down

8 How far away is the Vortex when it reaches point R?

The following information applies to questions 9–13.

In a shot-put event a 2.00 kg shot is launched from a height of 1.50 m, with an initial velocity of 8.00 m s^{-1} at an angle of 60.0° to the horizontal.

9 **a** What is the initial horizontal speed of the shot?

b What is the initial vertical speed of the shot?

c How long does it take the shot-put to reach its maximum height?

d What is the maximum height from the ground that is reached by the shot?

e How long after being thrown does the shot reach the ground?

f Calculate the total horizontal distance that the shot travels during its flight.

10 What is the speed of the shot when it reaches its maximum height?

11 What is the minimum kinetic energy of the shot during its flight?

12 What is the acceleration of the shot at its maximum height?

13 Which of the following angles of launch will result in the shot travelling the greatest horizontal distance before returning to its initial height?

A 15.0° **B** 30.0° **C** 45.0° **D** 60.0°

The following information applies to questions 14 and 15.

A student at the Australian Institute of Sport was able to establish that during its flight, a 2.00 kg shot experienced a force due to air resistance that was proportional to the square of its speed. The formula $F_a = 3.78 \times 10^{-5} v^2$ was determined, where F_a is the force due to air resistance and v is the instantaneous speed of the shot. The shot-put in one particular trial was launched from ground level at 7.50 m s^{-1} at an angle of 36.0° to the horizontal.

14 Calculate the maximum force due to air resistance that the shot experiences during its flight.

15 Calculate the value of the ratio of the forces acting on the shot as it is tossed:

$$\frac{F_g}{F_a\ (\text{max})}$$

What does your answer tell you about these forces?

***16** A ball of mass 200.0 g is dropped from a vertical height of 10.0 m onto a horizontal concrete floor. The ball rebounds upwards with an initial vertical velocity of 10.0 m s^{-1}. The time of interaction for this impact is 1.00 ms.

- **a** Calculate the momentum of the ball just before the impact.
- **b** What is the momentum of the ball just after the impact?
  - **A** 1000 kg m s^{-1} up
  - **B** 1.00 kg m s^{-1} up
  - **C** 2.00 kg m s^{-1} up
  - **D** 10.0 kg m s^{-1} up
- **c** What is the size of the impulse that has acted on the ball as it bounces?
  - **A** 2.80 N s
  - **B** 2.00 N s
  - **C** 0.800 N s
  - **D** 4.80 N s
- **d** Calculate the average net force that has acted on the ball.

The following information applies to questions 17 and 18.

A student supplies a constant force of 200.0 N at an angle of 60.0° to the horizontal to pull a 50.0 kg landing mat, initially at rest, across a horizontal gymnasium floor for 10.0 s. During this time a constant frictional force acts on the mat. The speed of the mat after 10.0 s is 3.00 m s^{-1}.

***17 a** What is the change in kinetic energy of the mat during this period?
- **b** How much work is done on the mat by the student?
- **c** How much energy is converted into heat during the 10.0 s interval?

***18 a** What is the power output of the student during the 10.0 s interval?
- **b** Calculate the power consumed by friction during this time.
- **c** What is the power output of the net force during this period?

The following information applies to questions 19–22.

A 50.0 kg boy stands on a 200.0 kg sled that is at rest on a frozen pond. The boy jumps off the sled with a velocity of 4.00 m s^{-1} east.

***19 a** What is the total momentum of the boy and the sled before he jumps off?
- **b** What is the momentum of the boy after he jumps?
- **c** What is the momentum of the sled after he has jumped?

After the boy has jumped off, he turns around and skates after the sled, jumping on with a horizontal velocity of 4.40 m s^{-1} west.

***20** Assuming the pond surface is frictionless, what is the velocity of the sled just before he jumps on?

***21** What is the speed of the boy once he is on the sled?

***22** As the boy jumps on the sled, what change in momentum is experienced by the:
- **a** sled?
- **b** boy?

The following information applies to questions 23–25.

Two air-track gliders, both travelling at 2.00 m s^{-1}, approach each other on an air-track. The gliders, with masses of 300.0 g and 100.0 g, are fitted with magnets so that the opposite poles are facing each other. The heavier glider is initially moving towards the east.

During the subsequent collision, the magnets stick together and the gliders move off with a common velocity.

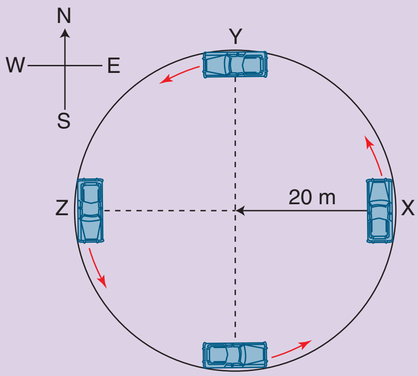
***23** Calculate their common velocity after colliding.

***24** What is the energy efficiency of this collision?
- **A** 100%
- **B** 0%
- **C** 25%
- **D** 75%

***25** What has happened to the 'missing' energy?
- **A** It has changed into momentum.
- **B** It has been transformed into heat and sound energy.
- **C** It has been stored as potential energy.
- **D** It has turned into magnetic energy.

The following information applies to questions 26–29.

A car of mass 1510 kg is driven at constant speed of 10.0 m s^{-1} around a level, circular roundabout. The centre of mass of the car is always 20.0 m from the centre of the track.



26 Use the answer key to answer this question.

Key:
- **A** 10 m s^{-1} north
- **B** 10 m s^{-1} east
- **C** 10 m s^{-1} south
- **D** 10 m s^{-1} west
- **E** None of these.

What is the velocity of the car at point:
- **a** X?
- **b** Y?
- **c** Z?

27 What is the period of revolution for this car?

28 a What is the centripetal acceleration of this car at point X?
- **b** Calculate the centripetal force acting on this car at point Y.
- **c** What is the unbalanced frictional force acting on the tyres at point Z?

29 Which one of the following statements is correct?
- **A** Since the car is moving with constant speed, there is zero net force acting on it.
- **B** The only force acting on the car is friction.
- **C** The frictional force between the tyres and the road provides the resultant force that keeps the car in its circular path.

30 The Ferris wheel at an amusement park has an arm radius of 10.0 m and its compartments move with a constant speed of 5.00 m s^{-1}.
- **a** Calculate the normal force that a 50.0 kg boy would experience from the seat when at the:
  - **i** top of the ride
  - **ii** bottom of the ride.
- **b** After getting off the ride, the boy remarks to a friend that he felt lighter than usual at the top of the ride. Which option explains why he might feel lighter at the top of the ride?
  - **A** He lost weight during the ride.
  - **B** The strength of the gravitational field was weaker at the top of the ride.
  - **C** The normal force there was larger than the gravitational force.
  - **D** The normal force there was smaller than the gravitational force.

2 Applying forces

Our Universe consists of perhaps 100 billion galaxies that are vast distances apart in space. It is expanding and this expansion is accelerating. The reason for this acceleration is currently a topic of major dispute among cosmologists.

The Universe is held together by the force of gravity. Isaac Newton adopted a theoretical approach in attempting to explain the behaviour of the heavenly bodies. He constructed an abstract framework of ideas to explain why things behaved the way they did. This was an advance on the earlier empirical approach of Johannes Kepler, who attempted to fit rules to match the data without trying to provide an explanation for these rules.

In 532 CE, Emperor Justinian commissioned the building of the Hagia Sophia in Constantinople (Istanbul in Turkey). This church consists of four piers on which four semicircular arches rest. The arches in turn support a large dome. The building still stands and is recognised as the finest example of Byzantine architecture in the world. At 72 m high, the Hagia Sophia was the tallest building in the world for about 1000 years!

This building is still standing despite being in an area that suffers from many earthquakes. There are few buildings of its age still standing in Istanbul and it has continued to stand while more modern buildings nearby have been destroyed by earthquakes. Now scientists think they know why.

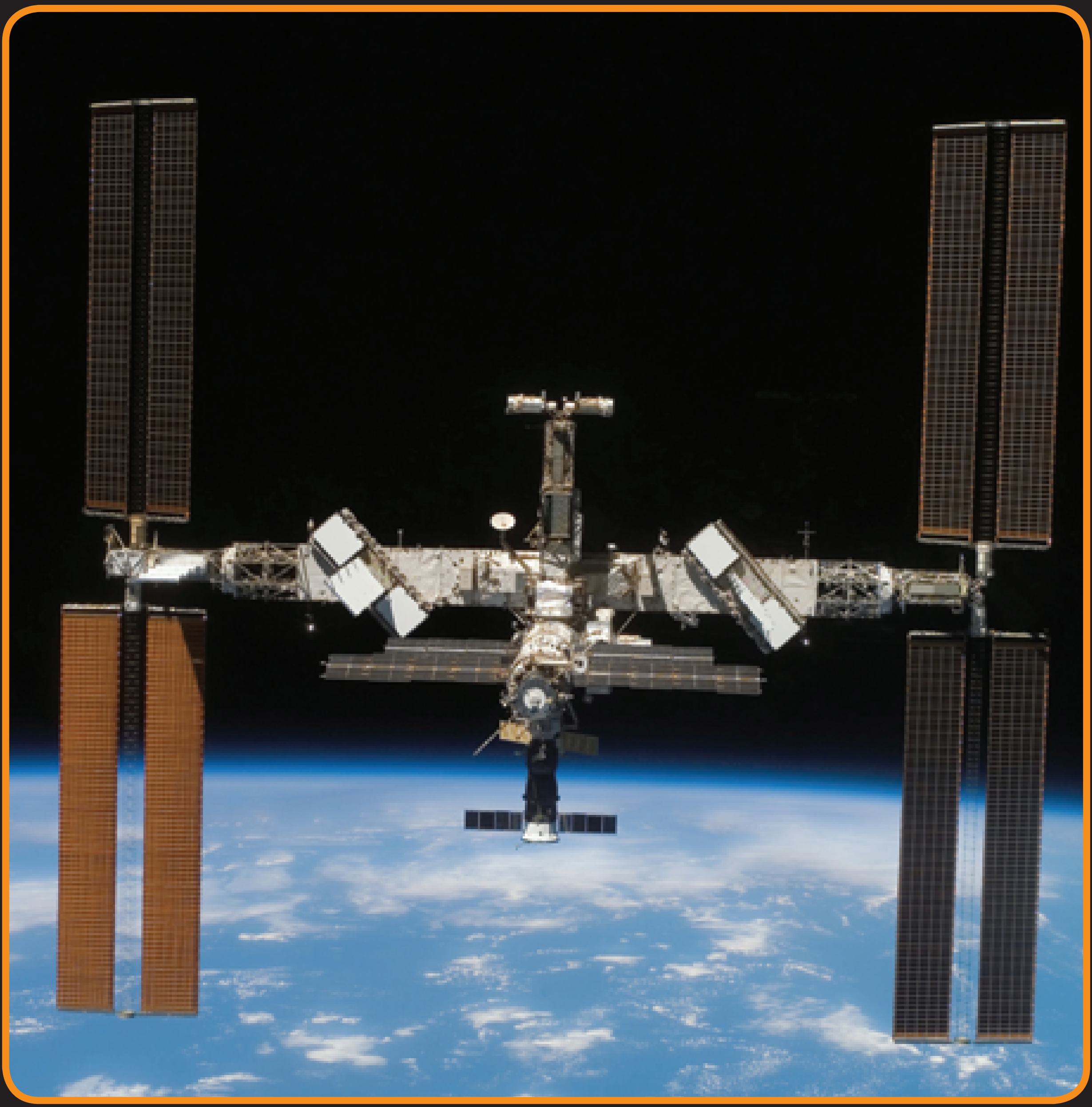
The architects in charge of its construction—Anthemius of Tralles and Isidore of Miletus—were highly skilled in mathematics and kinetics. They realised that making a rigid structure in such a location was bound to fail, so they deliberately constructed the Hagia Sophia so that it would flex and move a little. The cement that they used contained a calcium silicate matrix that is found in many modern types of cement. It is thought that they added volcanic ash to their mortar in order to give it energy-absorbing properties. It seems to have worked. The Hagia Sophia has withstood quakes that have measured up to 7.5 on the Richter scale.

In the design of buildings and other structures, engineers and architects must use their physics knowledge to determine the forces that act within the structures that they have created.

By the end of this chapter

you will have covered material from the study of applying forces including:

- Newton's law of universal gravitation
- gravitational fields
- satellite motion
- apparent weight, weightlessness and apparent weightlessness
- torque
- structures in equilibrium.



Physics file

Isaac Newton did not discover gravity—its effects have been known throughout human existence. But he was the first to understand the broader significance of gravity. Newton was supposedly sitting under an apple tree on his mother's farm at Woolsthorpe, in England, when an apple landed on his head. He looked up at the sky, noticed the Moon, and reasoned that the same force that made the apple fall to the ground also kept the Moon in its orbit about the Earth. The details of this story may or may not be true, but the way in which Newton developed his ideas about gravity is well documented.

2.1 Gravitational fields

Resource companies use sensitive gravity meters (gravimeters) in their search for mineral deposits. The gravitational field is weaker above a large deposit of oil than above surrounding areas because the oil has a lower density and is therefore less massive than the surrounding rock. Geologists detect this as they fly over the area in a survey plane. Ore bodies can be found in the same way. An ore body of minerals such as silver, lead or zinc can be almost twice as dense as the surrounding rock. In this case, the gravimeter would register a stronger gravitational field than above the surrounding area. The Pilbara iron ore deposit in Western Australia was identified from a gravity survey. BHP Billiton in Melbourne has developed a more sensitive device—an airborne gravity gradiometer—to help detect deposits containing diamonds.

From the laws of motion, Isaac Newton knew that a moving object would continue to move in a straight line with a constant speed unless an unbalanced force acted on it. He had also been thinking about the motion of the Moon and trying to work out why it moved in a circular orbit. The fact that it did not move in a straight line suggested to him that there must be a force acting on it. Newton proposed an idea that no-one else had even considered: that gravity from the Earth acted through space and exerted a force on the Moon.

He then generalised this idea and suggested that *gravitation* was a *force of attraction* that acted between *any* bodies. In this book we will use the convention of capital '*M*' and lower case '*m*' to represent a larger and smaller mass respectively.

The gravitational force acting between two bodies, *m* and *M*:

- is one of attraction, and can be considered to act from the centre of each mass
- acts equally and oppositely on each mass (Newton's third law): $\mathbf{F}_1 = -\mathbf{F}_2$
- is weaker if the masses are further apart. Gravitation acts in an inverse square manner, i.e. $F \propto \frac{1}{r^2}$; where *r* is the distance in metres between the *centres* of the masses
- depends directly on the mass of each body involved, i.e. $\mathbf{F} \propto m$ and $\mathbf{F} \propto M$.

Importantly, Newton showed that the force acts as if the mass of each body is located at its *centre of mass*. This is why the distance of separation *r* must be measured from the *centre* of each object.

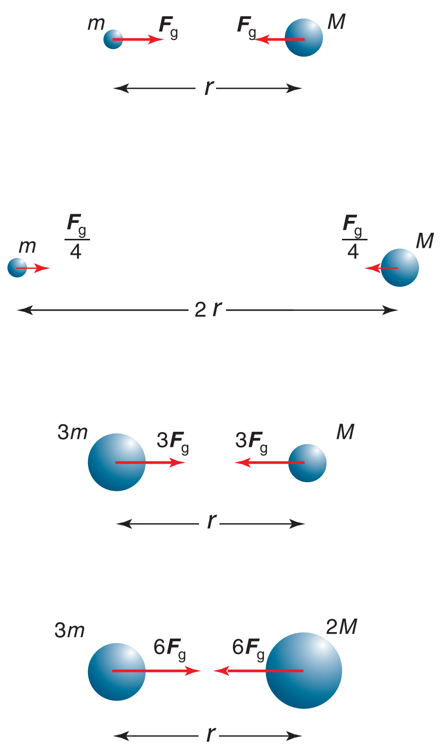


Figure 2.1
The gravitational force of attraction between two bodies depends on their separation and each of their masses.

The inclusion of a constant, *G*, gives ***NEWTON'S LAW OF UNIVERSAL GRAVITATION***:

$$\mathbf{F}_g = \frac{GMm}{r^2}$$

where $\mathbf{F}_g$ is the gravitational force acting on each body in newtons (N), *M* and *m* are the masses of the bodies in kilograms (kg) and *r* is the distance between the centres of the bodies in metres (m). The universal gravitational constant *G* is equal to 6.67×10^{-11} N m^2 kg^{-2}.

The extremely small value of the universal gravitational constant gives an indication of why gravitational forces are not noticeable between everyday objects. Consider two 1.00 kg dolls whose centres are 1.00 m apart (Figure 2.2).The gravitational force of attraction, F_g, between these bodies is given by:

$$F_g = \frac{GMm}{r^2}$$
$$= \frac{(6.67\times19^{-11})(1.00)(1.00)}{(1.00)^2}$$
$$= 6.67\times19^{-11}\ \text{N attraction}$$

This force is so small as to be insignificant. However, if one (or both) of the objects involved is *extremely massive*, then the gravitational force can have some effect. Consider the gravitational force of attraction between a 1.00 kg mass and the Earth. If the mass is at the Earth's surface, then the distance between it and the centre of the Earth is 6370 km or 6.37×10^6 m. Given that the mass of the Earth, M_E, is 5.98×10^{24} kg, the gravitational force of attraction that the 1.00 kg mass and the Earth exert on each other is:

$$F_g = \frac{GMm}{r^2}$$
$$= \frac{(6.67\times19^{-11})(5.98\times10^{24})(1.00)}{(6.37\times10^6)^2}$$
$$= 9.83\ \text{N attraction}$$

This value is the *weight* of the 1.00 kg mass. This weight force of 9.83 N acts on *both* the 1.00 kg mass and the Earth (Newton's third law). However, the effect of the force is vastly different. If the 1.00 kg mass is free to fall, it will accelerate at 9.83 m s^{-2} towards the Earth, whereas the Earth, with its enormous mass, will not be moved to any measurable degree by this 9.83 N force.

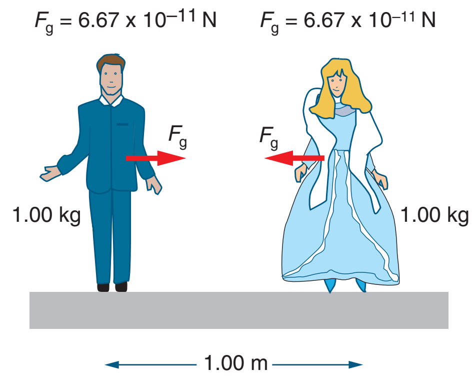


Figure 2.2
The gravitational force of attraction between these two dolls is so weak that it does not cause them to be thrown together.

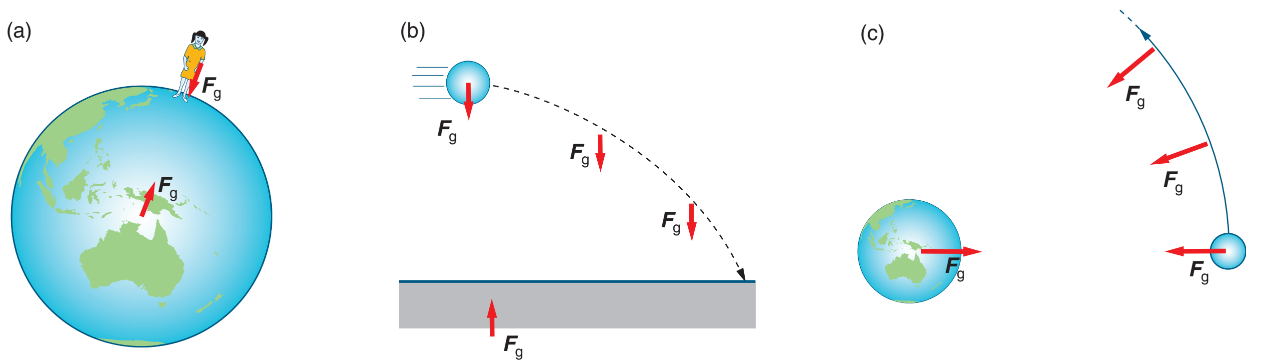


Figure 2.3
The effect of the gravitational force is significant when one of the masses involved is large. Gravitation is the force that (a) pulls you towards the Earth, (b) causes a projectile to fall towards the Earth, and (c) causes the Moon to 'fall' around the Earth.

Gravitation, although it is the weakest of the four fundamental forces that exist in nature (see Physics in action on page 58), is the most far-reaching of these forces; its influence stretches across the Universe.

Physics file

Newton explained the motion of the planets in terms of gravitational forces. He said that the planets were continually deviating from straight-line motion as a result of the gravitational force from the Sun that was continually acting on them. Einstein had a different explanation. His theory of general relativity proposed that the planets moved through space-time that had been warped by the Sun's gravitational field. As physicist John Wheeler noted, 'Matter tells space how to curve, and curved space tells matter how to move.'

Physics file

Newton's law of universal gravitation led to the discovery of Neptune in 1846. At that time, only the seven innermost planets out to Uranus were known. A French astronomer, Urbain Le Verrier, noticed a perturbation (a slight wobble) in the orbit of Uranus, and deduced that there must be another body beyond Uranus that was causing this to happen. Using just Newton's law of universal gravitation, he calculated the location of this body in the sky. Le Verrier informed the Berlin Observatory of his finding and, within half an hour of receiving his message, they had discovered Neptune.

✓ Worked Example 2.1A

A 10.0 kg watermelon falls a short distance to the ground. If the Earth has a radius of 6.37×10^6 m and a mass of 5.98×10^{24} kg, calculate the:

a gravitational force that the Earth exerts on the watermelon
b gravitational force that the watermelon exerts on the Earth
c acceleration of the watermelon towards the Earth
d acceleration of the Earth towards the watermelon.

Solution

a The gravitational force of attraction that the Earth exerts on the watermelon is:

$m_{wm} = 10.0$ kg

$M_E = 5.98 \times 10^{24}$ kg

$G = 6.67 \times 10^{-11}$ N m^2 kg^{-2}

$$F_{\text{E on wm}} = \frac{GMm}{r^2} = \frac{(6.67 \times 19^{-11})(5.98 \times 10^{24})(10.0)}{(6.37 \times 10^6)^2} = 98.3 \text{ N towards the Earth}$$

b The watermelon also exerts a force of attraction of 98.3 N towards the watermelon, on the Earth. These forces are an action/reaction pair. Even though the forces acting on the Earth and the watermelon are equal in size, the effect of these forces on each body is very different.

c The acceleration of the watermelon is:

$m_{wm} = 10.0$ kg

$F_{\text{E on wm}} = 98.3$ N

$$F_{\text{E on wm}} = ma_{wm}$$

$$a_{wm} = \frac{98.3}{10.0} = 9.83 \text{ m s}^{-2} \text{ towards the Earth}$$

This is, of course, very close to the average acceleration of all free-falling objects near the Earth's surface.

d The acceleration of the Earth is:

$m_{wm} = 5.98 \times 10^{24}$ kg

$F_{\text{wm on E}} = 98.3$ N

$$F_{\text{wm on E}} = ma_E$$

$$a_E = \frac{98.3}{5.98 \times 10^{24}} = 1.64 \times 10^{-23} \text{ m s}^{-2}$$

towards the watermelon

The motion of the Earth is hardly affected by the gravitational force of the watermelon.

✓ Worked Example 2.1B

The gravitational force that acts on a 1.20×10^3 kg space probe at the equator on the surface of Mars is 4.43×10^3 N. The radius of Mars is 3397 km at the equator. Without using the mass of Mars, determine the gravitational force that acts on the space probe when it is:

a 3397 km above the surface of Mars
b 6794 km above the surface of Mars.

Solution

a There is an inverse square relationship between gravitational force and the distance between the centres of the objects. Therefore, the product of the force and radius squared is a constant, so the force at any radius multiplied by that radius squared is equal to the force at any other radius multiplied by the radius squared:

$$F_{g1} \propto \frac{1}{r_1^2}$$

$$F_{g1} \times r_1^2 = \text{constant}$$

$$F_{g1} \times r_1^2 = F_{g2} \times r_2^2$$

When it is sitting on the surface of Mars, the probe is 3397 km from the centre of the planet and the gravitational force acting on it is 4.43×10^3 N. When the probe is at an altitude of 3397 km, it is 6794 km from the centre of Mars. This is double the distance from the centre compared to when it was on the surface. If the distance between the masses has doubled, then the size of the forces acting must be one-quarter of the original value:

$$F_{g1} = 4.43 \times 10^3 \text{ N} \qquad F_{g1} \times r_1^2 = F_{g2} \times r_2^2$$

$$r_1 = 3397 \times 10^3 \text{ m} \qquad F_{g2} = \frac{F_{g1} \times r_1^2}{r_2^2}$$

$$r_2 = 6794 \times 10^3 \text{ m} \qquad F_{g2} = \frac{(4.43 \times 10^3) \times (3397 \times 10^3)^2}{(6794 \times 10^3)^2}$$

$$= 1.11 \times 10^3 \text{ N towards Mars}$$

b At 6794 km above the surface of Mars, the probe is 10 191 km from the centre. This is three times its separation from the centre when it was on the surface, and so the gravitational force will be one-ninth of its original strength:

$$F_{g1} = 4.43 \times 10^3 \text{ N} \qquad F_{g1} \times r_1^2 = F_{g2} \times r_2^2$$

$$r_1 = 3397 \times 10^3 \text{ m} \qquad F_{g2} = \frac{F_{g1} \times r_1^2}{r_2^2}$$

$$r_2 = 10191 \times 10^3 \text{ m} \qquad F_{g2} = \frac{(4.43 \times 10^3) \times (3397 \times 10^3)^2}{(10191 \times 10^3)^2}$$

$$= 4.92 \times 10^2 \text{ N towards Mars}$$

Physics in action — How the mass of the Earth was first determined

The value of the universal gravitational constant, G, was first determined around 1844. The groundwork for this determination was done by Englishman Henry Cavendish in 1798. Cavendish, a rather eccentric man, was enormously wealthy and he devoted his life to scientific pursuits. He used the apparatus shown in Figure 2.4 to obtain the first direct measurement of the gravitational force of attraction between small objects. A large cage, Z, enclosed two pairs of lead spheres: a large pair, W, and a smaller pair, X, suspended from a metal rod that was itself suspended by a fine wire. The large spheres, initially close to the floor, were raised so that they were adjacent to but on opposite sides of the smaller spheres. The smaller spheres experienced a gravitational force of attraction towards the larger spheres. As they swung closer, the wire twisted. The amount of twist gave Cavendish a measure of the size of the gravitational forces. He needed a small telescope, T, illuminated by a lamp, L, to detect the tiny movement. These measurements enabled Cavendish to obtain the first calculation of the density of the Earth.

The mass of the Earth, M_E, could then be determined. The radius of the Earth, long known as 6.37×10^6 m, and the gravitational force on a 1.00 kg mass were used to calculate the Earth's mass:

$$F_g = \frac{GM_E m}{r^2}$$

$$M_E = \frac{F_g r^2}{Gm}$$

$$= \frac{(9.80)(6.37 \times 10^6)^2}{(6.67 \times 10^{-11})(1.00)}$$

$$= 5.96 \times 10^{-24} \text{ kg}$$

The value for G that was calculated later was only 1% different from the value that we use today.

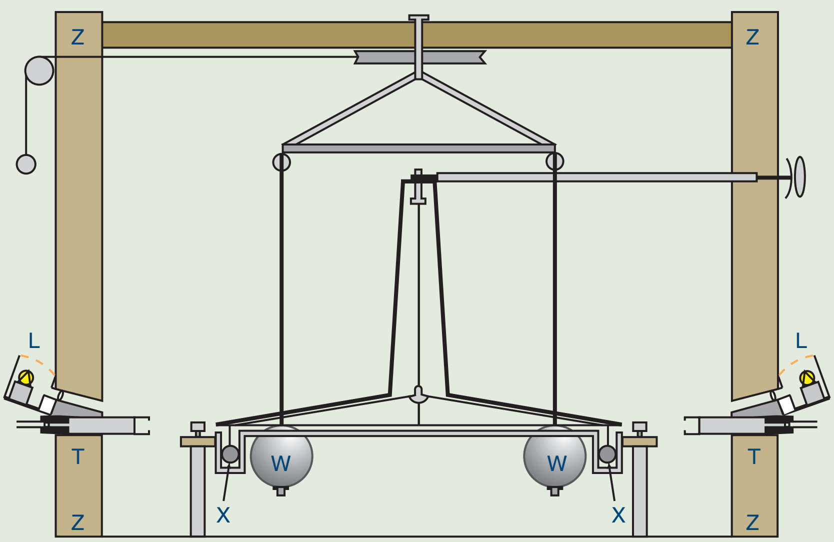


Figure 2.4
A diagram of the apparatus that was used by English physicist Henry Cavendish in 1798 to experimentally determine the value of the Earth's density

Physics in action — Gravity: the weakest force

Just four fundamental forces are responsible for the behaviour of matter, ranging from the smallest subatomic particles through to the most massive galaxies. These are the strong nuclear force, weak nuclear force, electromagnetic force and gravitational force.

The strongest of these forces, the ***strong force***, is the force that binds the nucleus. It is a force of attraction that acts between nucleons (protons and neutrons), and is strong enough to overcome the repulsion between the protons in the nucleus. It only acts over a very short distance, of the order of 10^{-15} m, which is approximately the distance between adjacent nucleons.

The ***weak force*** is an extremely short-range force that acts inside atomic nuclei. It is responsible for radioactive processes such as beta decay. The weak force is much weaker than the strong nuclear and electromagnetic forces.

Electromagnetic forces act between charged particles: like charges repel and unlike charges attract. Compared with the strong force, the electromagnetic force is a long-range force. It acts to hold electrons in their orbits around atomic nuclei, but its strength varies as the inverse square of the distance between the charges. The electromagnetic force can act over an infinite distance.

The ***gravitational force*** is also a long-range force that acts between all bodies. Gravitational forces not only make things fall to the ground when they are dropped, they also extend across space and hold the planets in their orbits around the Sun—they hold galaxies together. The gravitational force affects the rate at which the Universe is expanding.

Physicists are currently scanning the Universe in an effort to find gravity waves. Some predict that collapsing stars will create ripples in space-time that we will be able to detect.

While gravitation, along with electromagnetism, has the longest range of the four fundamental forces, it is by far the weakest of these forces. Its strength is just 10^{-25} times that of even the weak force. Over extremely small distances, the strong, electromagnetic and weak forces are dominant and the gravitational attraction between the particles is of no significance. But on the much larger scale of outer space, the opposite situation arises: gravitation rules! The electromagnetic force has no influence over these distances because, at a distance, matter appears uncharged.

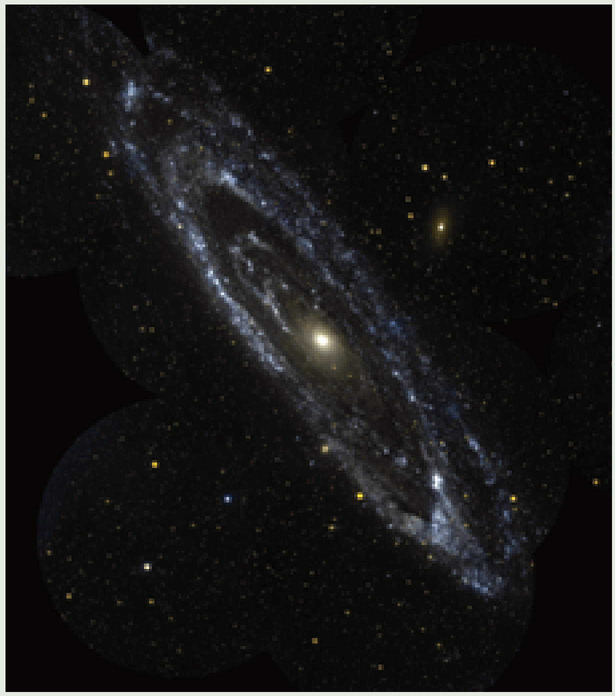


Figure 2.5
Gravitational forces hold planets in their orbits around the Sun.

Gravitational fields

In physics, a *field* is a region in which an object experiences a force. For example, around a magnet there is a magnetic field, ***B***. This field affects objects with magnetic properties, such as iron and cobalt. They will experience a force of attraction towards the magnet if they are placed in this region.

Around electrically charged objects, there is an electric field, ***E***. If a charged particle is placed in this region, it will experience an electric force that can be either an attraction or a repulsion, depending on the sign of the charges. A *gravitational field*, ***g***, is a region around a *mass* in which other masses will experience a *gravitational force*.

Using diagrams to represent gravitational fields

The Earth has a gravitational field around it. A mass that is close to the Earth experiences a force of attraction towards it. Gravitational forces are *always* forces of attraction.

The gravitational field in your classroom can be analysed by examining the gravitational force that acts on a sample mass at different points in the room. If you take a 1.00 kg mass to various parts of the room and use a force-meter to measure the gravitational force that acts on it, you will find that the gravitational force is approximately 9.80 N vertically downwards at every point in the room.

This experiment shows that the strength of the gravitational field in the room is *uniform*. A gravitational field, ***g***, can be represented diagrammatically by lines with arrows that show the direction of the force on the object. The gravitational field in your classroom is uniform, so the field is represented by evenly spaced parallel lines, in the direction of the force.

Now imagine that you have a giant ladder that stretches up into space. If you could take the 1.00 kg mass and force-meter and repeat this exercise at higher and higher altitudes, you would find that the gravitational force on the mass becomes less and less. In other words, the force that the 1.00 kg mass pushes down on the force-meter (its weight) decreases as you move further from the centre of the Earth.

At the Earth's surface, a force of 9.80 N acts on the mass. At an altitude of 1000 km, the force-meter would read 7.30 N, and at 5000 km it would show just 3.10 N. If this exercise were to be repeated at different places around the world, the results would be the same.

On a larger scale, the gravitational field around the Earth is directed towards the centre of the Earth and is not uniform. It becomes weaker at higher altitudes.

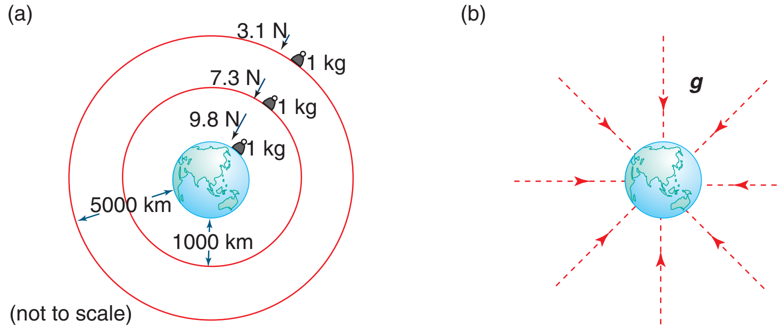


Figure 2.7
(a) The gravitational force acting on a 1.00 kg mass becomes smaller at greater distances from the Earth. (b) This non-uniform gravitational field, ***g***, is represented by radial lines directed towards the centre of the Earth. The field is strongest where the lines are closest.

The field lines in Figure 2.7b are not equally spaced and this indicates that the gravitational field strength is not uniform. Close to the Earth where the field lines are close together, the gravitational field is relatively strong, but at higher altitudes where the field lines are spread out, the gravitational field is weaker, tending to zero at a large distance from the Earth.

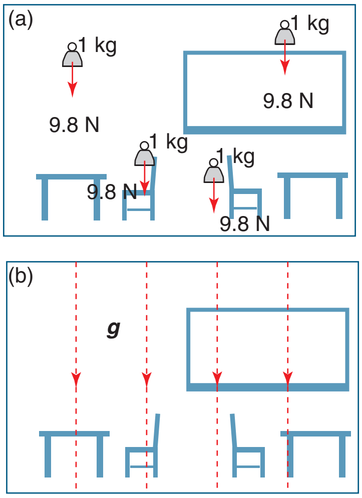


Figure 2.6
(a) The gravitational force on a 1.00 kg mass is constant at 9.80 N everywhere in your classroom. (b) The uniform gravitational field *g* is represented by evenly spaced parallel lines in the direction of the force.

Physics file

In this textbook a distinction is made between the force on an object due to gravity and the weight force of an object. If the object is in a gravitational field, then a gravitational force is acting on it due to the attraction of its mass to the mass of the object creating the field. If the object is in contact with a surface, then this gravitational attraction causes the object to push down on the surface—this is the weight force. The distinction is helpful when we consider an object in free-fall; it has a gravitational force acting on it, but it is weightless until it lands on a surface. A lack of understanding about this simple distinction has caused misconceptions such as there being no gravity in space, as astronauts are described as weightless. Of course we know that gravity extends far into space, because if there was no gravity in space planets would not revolve around the Sun!

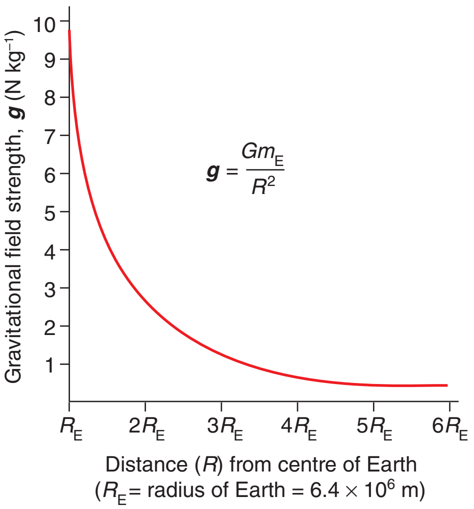


Figure 2.8

The gravitational field strength around the Earth. There is an inverse square relationship between the field strength, $\boldsymbol{g}$, and the distance, R, from the centre of the Earth.

Calculating gravitational field strength

The magnitude of a gravitational field is known as the *gravitational field strength*, $\boldsymbol{g}$. This is defined as the gravitational force that acts on each kilogram of a body in the field. Since gravitational field strength, $\boldsymbol{g}$, is defined as gravitational force per unit mass, we can say that:

$$\boldsymbol{g} = \frac{\boldsymbol{F}_g}{m} = \frac{GMm}{r^2 m} = \frac{GM}{r^2}$$

The ***GRAVITATIONAL FIELD STRENGTH*** is given by:

$$\boldsymbol{g} = \frac{GM}{r^2}$$

where $\boldsymbol{g}$ is the gravitational field strength in newtons per kilogram (N kg^{-1}), which can also be shown to be (m s^{-2}), G is the universal gravitation constant (6.67×10^{-11} N m^2 kg^{-2}), M is the mass of the object around which the gravitational field spreads (kg) and r is the distance from the centre of mass of the object to the point at which the field strength is to be determined (m).

If $\boldsymbol{g}$ is known, the force due to gravity, $\boldsymbol{F}_g$, of a mass in the field can be found by using $\boldsymbol{F}_g = m\boldsymbol{g}$.

The ***WEIGHT*** of a body is given by:

$$\boldsymbol{F}_{wt} = m\boldsymbol{g}$$

where m is the mass of an object, in the gravitational field and on a surface (kg), and $\boldsymbol{g}$ is the gravitational field strength at that surface (N kg^{-1}),

The gravitational field strength of the Earth is approximately 9.80 N kg^{-1}, but only at the surface of the Earth. Nevertheless, for problems involving the motion of falling bodies and projectiles *close to the Earth's surface,* it is reasonable to assume that the gravitational field strength is *constant* at 9.80 N kg^{-1}. However, if an object falls from, or is launched to, a high altitude, the *changing* gravitational field strength should be taken into account.

table 2.1 Gravitational fields around the Solar System

Object	Gravitational field strength at surface (N kg^{-1})
Sun*	270
Mercury	3.3
Venus	8.1
Earth	9.8
Mars	3.6
Ceres	0.27
Jupiter*	24.6
Saturn*	10.4
Uranus*	8.2
Neptune*	11.2
Pluto	0.60
Eris	0.68

*These objects are gaseous and do not have solid surfaces.

✓ Worked Example 2.1C

a Calculate the gravitational field strength, $\boldsymbol{g}$, at the surface of the Earth (mass of Earth is 5.98×10^{24} kg, radius of Earth is 6.37×10^{6} m).
b Calculate the gravitational field strength, $\boldsymbol{g}$, at the surface of the Moon (mass of Moon is 7.35×10^{22} kg, radius of Moon is 1.74×10^{6} m).
c Determine the weight of an astronaut, whose total mass is 100.0 kg, at each of these locations.
d Determine your own weight on the surface of the Earth and on the surface of the Moon.

Solution

a At the Earth's surface:

$M_E = 5.98 \times 10^{24}$ kg

$G = 6.67 \times 10^{-11}$ N m^2 kg^{-2}

$R = 6.37 \times 10^{6}$ m

$$\boldsymbol{g} = \frac{GM}{R^2} = \frac{(6.67 \times 10^{-11})(5.98 \times 10^{24})}{(6.37 \times 10^{6})^2} = 9.82 \text{ N kg}^{-1}$$

b At the Moon's surface:

$M_m = 7.35 \times 10^{22}$ kg

$G = 6.67 \times 10^{-11}$ N m^2 kg^{-2}

$R = 1.74 \times 10^{6}$ m

$$\boldsymbol{g} = \frac{GM}{R^2} = \frac{(6.67 \times 10^{-11})(7.35 \times 10^{22})}{(1.74 \times 10^{6})^2} = 1.62 \text{ N kg}^{-1}$$

c On Earth's surface, the astronaut would weigh:

$m = 100.0$ kg

$\boldsymbol{g}_E = 9.82$ N kg^{-1}

$$\boldsymbol{F}_{wt} = m\boldsymbol{g} = (100.0)(9.82) = 9.82 \times 10^{2} \text{ N}$$

On the Moon, the astronaut would weigh only:

$m = 100.0$ kg

$\boldsymbol{g}_m = 1.62$ N kg^{-1}

$$\boldsymbol{F}_{wt} = m\boldsymbol{g} = (100.0)(1.62) = 1.62 \times 10^{2} \text{ N}$$

d On Earth's surface, you would substitute your own mass into the first part of question (c) above, and on the surface of the Moon you would substitute your mass in the second calculation in part (c). You will need to calculate this for yourself.

Practical activity
6 Acceleration down an incline

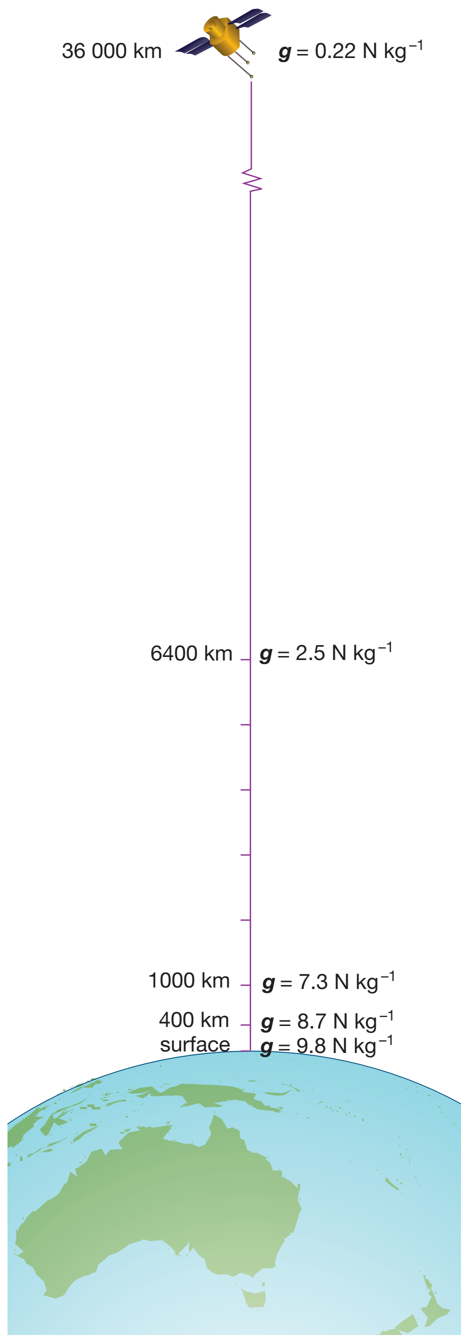


Figure 2.9
At 400 km above the Earth's surface, the gravitational field strength is 8.70 N kg^{-1}. The International Space Station has been in orbit at nearly this altitude since 1998. Australia's Optus communications satellites are in orbit at an altitude of about 36 000 km where Earth's gravitational field strength is only 0.22 N kg^{-1}.

2.1 SUMMARY Gravitational fields

- Gravitation is a force of attraction that acts between all bodies.
- The gravitational force acts equally, but oppositely, on each of the two bodies.
- The gravitational force is directly related to the masses: $F_g \propto M, m$.
- The gravitational force is weaker when the bodies are further apart. There is an inverse square relationship between the force and the distance of separation of the bodies:

$$F_g \propto \frac{1}{r^2}$$

- The separation distance of the bodies is measured from the centre of mass of each object.
- Newton's law of universal gravitation is:

$$F_g = \frac{GMm}{r^2}$$

- The constant of universal gravitation is $G = 6.67 \times 10^{-11}$ N m^2 kg^{-2}
- A gravitational field is a region in which any object with mass will experience a force.
- The gravitational field strength, $\boldsymbol{g}$, is greater around more massive central objects, and becomes weaker at greater distances from the central body.
- Gravitational field strength, $\boldsymbol{g}$, is given by:

$$\boldsymbol{g} = \frac{GM}{r^2}$$

- At the surface of the Earth, the gravitational field strength is 9.8 N kg^{-1}.
- The gravitational force acting on an object that is on a surface is called the weight, F_{wt} of the object and is given by:

$$F_{wt} = mg$$

2.1 Questions

Assume that $G = 6.67 \times 10^{-11}$ N m^2 kg^{-2}, and that distances are measured between the centres of the bodies.

1 For this question, assume that:

mass of Earth = 5.98×10^{24} kg; radius of Earth = 6.37×10^6 m; mass of the Moon = 7.35×10^{22} kg; mean radius of Moon's orbit = 3.84×10^8 m.

Calculate the gravitational force of attraction that exists between the following objects.

a A 100.0 g apple and a 200.0 g orange that are 50.0 cm apart

b The Earth and a satellite of mass 2.00×10^4 kg in orbit at an altitude of 600.0 km

c The Moon and the Earth

d A proton of mass 1.67×10^{-27} kg and an electron of mass 9.11×10^{-31} kg separated by 5.30×10^{-11} m in a hydrogen atom

2 An astronaut standing on the surface of the Moon experiences a gravitational force of attraction of 160 N. He then moves away from the surface of the Moon to an altitude where the gravitational force is 40 N.

a How far from the centre of the Moon is this new location in terms of the radius of the Moon?

b The astronaut now travels to another location at a height of three Moon radii above the surface. Calculate the gravitational force at this altitude.

3 Mars has two natural satellites. Phobos has a mean orbital radius of 9.4×10^6 m, while the other moon, Deimos, is located at a mean distance of 2.35×10^7 m from the centre of Mars. Data: mass of Mars = 6.42×10^{23} kg; mass of Phobos = 1.08×10^{16} kg; mass of Deimos = 1.8×10^{15} kg.

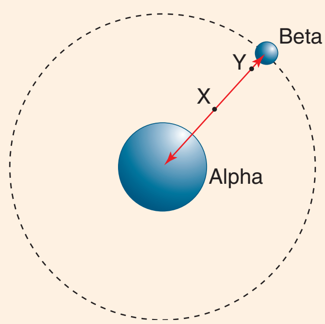
Calculate the value of the ratio:

$$\frac{\text{gravitational force exerted by Mars on Phobos}}{\text{gravitational force exerted by Mars on Deimos}}$$

4 An astronaut travels away from Earth to a region in space where the gravitational force due to Earth is only 1.00% of that at Earth's surface. What distance, in Earth radii, is the astronaut from the centre of the Earth?

The following information applies to questions 7 and 8.

The planet Alpha, whose mass is M, has one moon Beta of mass $0.01M$. The mean distance between the centres of Alpha and Beta is R.



5 a If an asteroid is at point X, exactly halfway between the centres of Alpha and Beta, calculate the value of the ratio:

$$\frac{\text{force exerted on asteroid by Alpha}}{\text{force exerted on asteroid by Beta}}$$

b At what distance, expressed in terms of R, from the planet Alpha, along a straight line joining the centres of Alpha and Beta, will the ratio expressed in part a be equal to 8100?

6 Point Y represents the distance from planet Alpha where the magnitude of the net gravitational force is zero. What is this distance in terms of R?

The following information applies to questions 7–9.

The masses and radii of three planets are given in the following table.

Planet	Mass (kg)	Radius (m)
Mercury	3.30×10^{23}	2.44×10^{6}
Saturn	5.69×10^{26}	6.03×10^{7}
Jupiter	1.90×10^{27}	7.15×10^{7}

7 Calculate the gravitational field strength, $\boldsymbol{g}$, at the surface of each planet.

8 Using your answers to Question 7, calculate the weight of an 80.0 kg astronaut on the surface of:
- a Mercury
- b Saturn
- c Jupiter.

9 The result of your calculation for Question 7 should indicate that the gravitational field strength for Saturn is very close in value to that on Earth, i.e. approximately 9.80 N kg^{-1}. However, the Earth's radius and mass are very different from those of Saturn. How do you account for the fact that the two planets have similar gravitational field strengths?

10 There is a point between the Earth and the Moon where the total gravitational field is zero. The significance of this is that returning lunar missions are able to return to Earth under the influence of the Earth's field once they pass this point. Given that the mass of Earth is 5.98×10^{24} kg, the mass of the Moon is 7.35×10^{22} kg and the radius of the Moon's orbit is 3.84×10^{8} m, calculate the distance of this point from the centre of the Earth.

2.2 Satellite motion

Artificial satellites

Since the Space Age began in the 1950s, thousands of artificial satellites have been launched into orbit around the Earth. These are used for a multitude of different purposes including communications. Our deep-space weather pictures come from the Japanese MTSAT-1R satellite, while close-range images are received from the American NOAA satellites. The Hubble Space Telescope is used to view objects right at the edge of the Universe. Other satellites are helping to geologically map the surface of the Earth, to monitor the composition of the atmosphere, to determine the extent of rainforest destruction and monitor the extent of global warming and climate change.

Artificial and natural satellites are not propelled by rockets or engines. They orbit in *free-fall* and the only force acting is the gravitational attraction between themselves and the body about which they orbit. Artificial satellites are often equipped with tanks of propellant that are squirted in the appropriate direction when the orbit of the satellite needs to be adjusted.

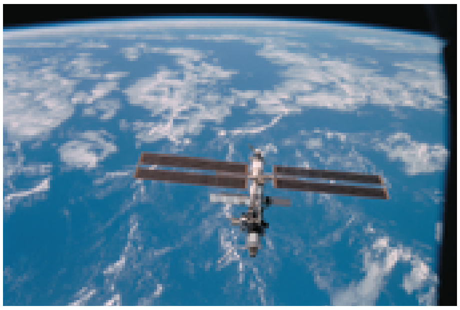


Figure 2.10
Australia's Optus D1 and D2 communications satellites were released from the cargo bays of space shuttles, several hundred kilometres above the Earth's surface, in 2006 and 2007, respectively. The propulsion system on the satellites then moved them to their geostationary orbits at an altitude of 36 000 km.

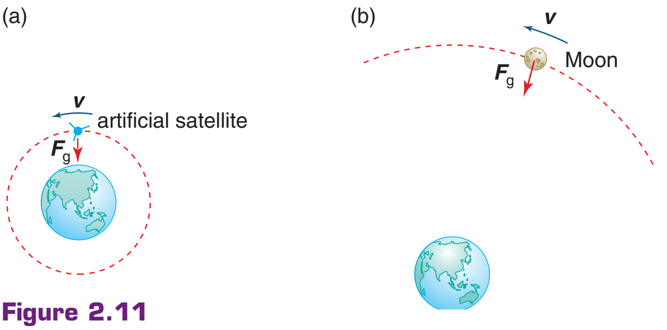


Figure 2.11
The only force acting on (a) an artificial satellite and (b) a natural satellite is the gravitational attraction of the central body.

Physics file

The first artificial satellite, Sputnik 1, was launched by the Soviet Union on 4 October 1957. It was a metal sphere just 58 cm in diameter and 84 kg in mass. Sputnik 1 orbited at an altitude of 900 km and had an orbital period of 96 minutes. It carried two radio transmitters that emitted a continuous series of beeps that were picked up by amateur radio operators around the world.

Physics file

The first space probe to leave the Solar System was Pioneer 10. It was launched in 1972 and passed the orbit of Pluto in 1984. Radio signals from Pioneer 10 finally cut out in 2003. It is estimated that Pioneer 10 will travel through interstellar space for another 80 000 years before encountering another star.

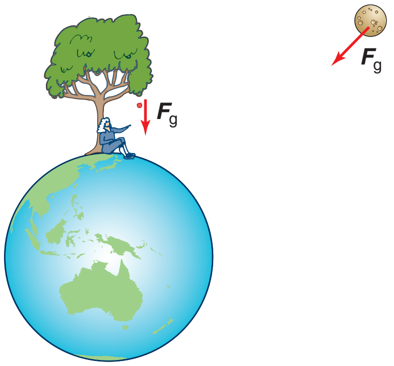


Figure 2.12
Newton realised that the gravitational attraction of the Earth was determining the motion of both the Moon and the apple.

Figure 2.13
Newton's original sketch shows how a projectile that was fired fast enough would fall all the way around the Earth and become an Earth satellite.

A *satellite* is an object in a *stable orbit* around another object. Isaac Newton developed the notion of satellite motion while working on his theory of gravitation. He was comparing the motion of the Moon with the motion of a falling apple and realised that it was the gravitational force of attraction towards the Earth that determined the motion of both objects. He reasoned that if this force of gravity was not acting on the Moon, it would move with constant velocity in a straight line at a tangent to its orbit.

Newton proposed that the Moon, like the apple, was also falling. It was continuously falling to the Earth without actually getting any closer to the Earth. He devised a thought experiment in which he compared the motion of the Moon with the motion of a cannonball fired horizontally from the top of a high mountain.

In this thought experiment, if the cannonball was fired at a low speed, it would not travel a great distance before gravity pulled it to the ground. If it was fired with a greater velocity, it would follow a less curved path and land a greater distance from the mountain. Newton reasoned that, if air resistance was ignored and if the cannonball was fired fast enough, it could travel around the Earth and reach the place from where it had been launched. At this speed, it would continue to circle the Earth indefinitely. In reality, satellites could not orbit the Earth at low altitudes, because of air resistance. Nevertheless, Newton had proposed the notion of an artificial satellite almost 300 years before one was actually launched.

Natural satellites

There are many different *natural satellites*. The planets and the asteroids are natural satellites of the Sun.

The Earth has one natural satellite—the Moon. The largest planets—Jupiter, Saturn and Uranus—have many natural satellites in orbit around them with more being discovered every year. The number of planetary moons is shown in Table 2.2.

table 2.2 The moons of the planets in the Solar System. A moon is a natural satellite

Planet	Number of moons
Mercury	0
Venus	0
Earth	1
Mars	2
Jupiter	>60
Saturn	>60
Uranus	27
Neptune	13

Satellites in circular orbits

The *gravitational* force that acts on a satellite is always directed *towards the centre of the central mass*, so the centre of the orbit must be the centre of the central mass. If the gravitational force that acts on the satellite is always *perpendicular* to the velocity of the satellite, the orbit is circular, as discussed in Chapter 1. The gravitational force does not cause the satellite to speed up or slow down; it only acts to change its direction of motion. The satellite will continually fall towards the Earth, and it will fall at the same rate at which the curved surface of the Earth is falling away from it, and so the satellite never actually gets any closer to the Earth as it orbits.

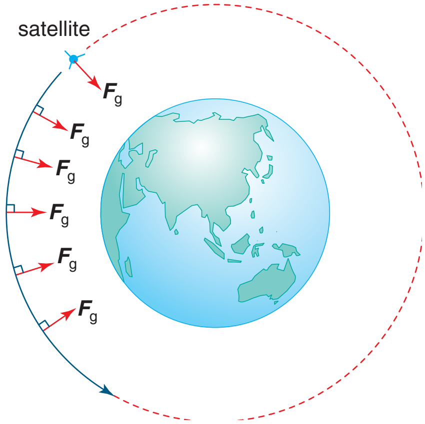


Figure 2.14
The gravitational force that acts on a satellite in a circular orbit is always at right angles to the velocity of the satellite. The force is directed towards the centre of the central mass and gives the satellite a centripetal acceleration.

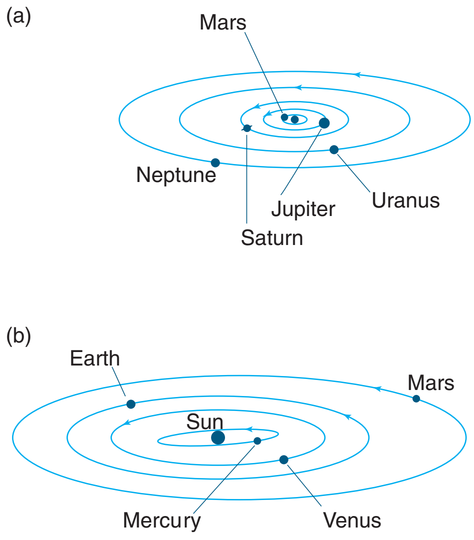


Figure 2.15
The outer planets of the Solar System have such large orbits that they cannot be shown on the same scale diagram as the inner planets. (a) The outer, or Jovian, planets are spread apart. If you travelled from the Sun to Neptune, by the time you reached Uranus you would be just one-sixth of the way. (b) The inner, or terrestrial, planets are relatively small and close together. The asteroid belt lies between the orbits of Mars and Jupiter.

As you will recall from section 2.1, any object that is falling freely in a gravitational field will have an *acceleration* that is determined by the *gravitational field strength*. A satellite is in *free-fall*. The only force acting on it is the gravitational attraction of the central body. This provides the necessary centripetal force, and causes the satellite to move in a circular path with constant speed. The size of the centripetal acceleration is determined by the gravitational field strength at the location. For example, the Hubble Space Telescope (HST) is in orbit at an altitude of 600 km. At this location, the gravitational field strength is 8.20 N kg^{-1}. The circular motion of the HST means that it has a centripetal acceleration of 8.20 m s^{-2} as it orbits.

The European Space Agency (ESA) is establishing its own satellite positioning system, to be known as Galileo. This will consist of 30 satellites and is expected to be operating by 2011.

Physics file

The images found using Google Earth and Google Maps are not in real time. The images are usually 1–3 years old and are a combination of satellite images and aerial photography.

table 2.3 Data for three artificial Earth satellites

Satellite	Altitude (km)	Orbital radius (km)	g (N kg^{-1})	Period	Speed (km s^{-1})	Acceleration (m s^{-2})
ISS	380	6 760	8.80	92 min	7.7	8.80
NAVSTAR GPS	20 200	26 500	0.57	12 h	3.9	0.57
Optus D2	35 900	42 300	0.22	24 h	3.1	0.22

It can be seen from Table 2.3 that the acceleration of each satellite is determined by the gravitational field strength in its particular orbit. Higher altitude orbits have weaker gravitational fields. Therefore, the gravitational forces acting on satellites are also weaker, and so their rates of acceleration towards the central body are smaller than those of low-orbit satellites.

Calculating the orbital properties of a satellite

For a satellite in a stable *circular orbit* of radius r and period T, equations that were used to analyse circular motion in Chapter 1 can be used again. The speed, v, of the satellite can be calculated from its motion for one

Physics file

There are 31 NAVSTAR satellites in orbit. These satellites form the global positioning system (GPS) network. GPS receiving units have a wide range of commercial and military applications. They allow people to know their location on Earth to within a few metres. GPS units help save bushwalkers who are lost and allow weekend sailors to know where they are. They are also used for the satellite navigation systems in many cars.

revolution. It will travel a distance equal to the circumference of the circular orbit ($2\pi r$) in a time period T.

The speed of a satellite is given by:

$$v = \frac{2\pi r}{T}$$

The centripetal acceleration, $\boldsymbol{a}_c$, of the satellite can also be calculated by considering its circular motion, or by determining the gravitational field strength at its location.

The speed relationship for circular motion can be substituted into the centripetal acceleration formula to give:

$$a_c = \frac{v^2}{r} = \frac{4\pi^2 r}{T^2}$$

Given that the centripetal acceleration of the satellite is equal to the gravitational field strength at the location of its orbit, it follows that:

SATELLITE ACCELERATION is given by:

$$\boldsymbol{a}_c = \frac{v^2}{r} = \frac{4\pi^2 r}{T^2} = \frac{GM}{r^2} = \boldsymbol{g}$$

where $\boldsymbol{a}_c$ is the centripetal acceleration of the satellite (m s^{-2}), v is the orbiting speed (m s^{-1}), r is the distance between the centres of mass of the central body and the orbiting object (m), T is the time period for one revolution of the orbiting object (s), G is the universal gravitational constant (6.67×10^{-11} N m^2 kg^{-2}), M is the mass of the central object around which the satellite is orbiting (kg) and $\boldsymbol{g}$ is the gravitational field strength at the distance r (N kg^{-1}).

The gravitational force, $\boldsymbol{F}_g$, acting on the satellite can then be found by using Newton's second law.

GRAVITATIONAL FORCE is given by:

$$\boldsymbol{F}_c = \frac{mv^2}{r} = \frac{4\pi^2 rm}{T^2} = \frac{GMm}{r^2} = m\boldsymbol{g} = \boldsymbol{F}_g$$

where $\boldsymbol{F}_c$ is the centripetal force acting on the satellite (N), m is the mass of the satellite (kg), v is the orbiting speed (m s^{-1}), r is the distance between the centres of mass of the central body and the orbiting object (m), $\boldsymbol{T}$ is the time period for one revolution of the orbiting object (s), G is the universal gravitational constant (6.67×10^{-11} N m^2 kg^{-2}), M is the mass of the central object around which the satellite is orbiting (kg), $\boldsymbol{g}$ is the gravitational field strength at the distance r (N kg^{-1}) and $\boldsymbol{F}_g$ is the gravitational force (N).

These relationships can be manipulated to determine any feature of a satellite's motion—its *speed, radius* of orbit or *period* of orbit. They can also be used to find the *mass* of the central body. As with the motion of free-falling objects at the Earth's surface, the *mass* of the satellite itself has *no bearing* on any of these quantities.

Physics file

Communications satellites are most useful if they are accessible 24 hours a day. Low-orbit satellites are of limited use in this regard because they may complete up to 15 orbits each day and so will be over their home country for only a small portion of the day. Geostationary or geosynchronous satellites are required. These are satellites that orbit at the same rate at which the Earth spins; they have a period of 24 hours. This is achieved by placing the satellites in an orbit above the Equator at a radius 42 300 km, approximately 36 000 km above the surface of the Earth. They then turn with the Earth and so remain fixed above one point on the Equator. The position of these satellites is indicated by the angle and direction of satellite dishes. In Australia, they point in a northerly direction to a position in the sky above the Equator. In countries on the Equator, satellite dishes often point straight up!

Physics file

Edwin Hubble was born in Missouri, USA, and was the third of eight children. He was a talented athlete and a gifted scholar, winning a Rhodes scholarship to Oxford University, where he studied law. After graduating and setting up a law practice in Kentucky, he realised that he was more interested in physics and astronomy. At the time, it was thought that the Milky Way was the entire Universe. Hubble, using the Mt Wilson telescope, was able to show that the Milky Way was just one galaxy out of millions of galaxies throughout the Universe. He also showed that these galaxies were moving apart from each other and that the Universe was expanding. This allowed the age of the Universe to be determined and supported the Big Bang theory that explained the origins of the Universe.

Figure 2.16
Edwin Hubble (1889–1953), after whom the Hubble Space Telescope was named.

✓ Worked Example 2.2A

Optus D2 is a geostationary satellite. Its period of orbit is 24.0 hours, so that it revolves at the same rate at which the Earth turns. Given that the mass of the Earth is 5.98×10^{24} kg and the mass of Optus D2 is 1160 kg, calculate:

a Optus D2's orbital radius
b the gravitational field strength at this radius
c Optus D2's orbital speed
d Optus D2's acceleration.

Solution

a $T = 24.0 \text{ h} = 8.64 \times 10^4 \text{ s}$

$G = 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$

$M_E = 5.98 \times 10^{24} \text{ kg}$

$$F_c = F_g$$

$$\frac{m4\pi^2 r}{T^2} = \frac{GMm}{r^2}$$

$$r^3 = \frac{GMT^2}{4\pi^2}$$

$$r = \sqrt[3]{\frac{(6.67 \times 10^{-11})(5.98 \times 10^{24})(8.64 \times 10^4)^2}{4\pi^2}}$$

$$= 4.23 \times 10^7 \text{ m}$$

So the orbital radius of Optus D2 is 42 300 km. This is about 5.5 Earth radii from the surface of the Earth.

b $G = 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$

$M_E = 5.98 \times 10^{24} \text{ kg}$

$r = 4.23 \times 10^7 \text{ m}$

$$\boldsymbol{g} = \frac{GM}{r^2}$$

$$= \frac{(6.67 \times 10^{-11})(5.98 \times 10^{24})}{(4.23 \times 10^7)^2}$$

$$= 0.223 \text{ N kg}^{-1}$$

This satellite is in orbit at a great distance from the Earth and the gravitational field strength at its location is just 0.223 N kg^{-1}.

c The speed of Optus D2 is given by:

$T = 24.0 \text{ h} = 8.64 \times 10^4 \text{ s}$

$r = 4.23 \times 10^7 \text{ m}$

$$v = \frac{2\pi r}{T}$$

$$= \frac{2\pi(4.23 \times 10^7)}{8.64 \times 10^4}$$

$$= 3.07 \times 10^3 \text{ m s}^{-1}$$

i.e. about 3 km s^{-1}

d The acceleration of the satellite as it orbits is equal to the gravitational field strength at this radius as calculated in part b, i.e. $\boldsymbol{a} = \boldsymbol{g} = 0.223 \text{ m s}^{-2}$ towards the centre of the Earth.

✓ Worked Example 2.2B

Ganymede is the largest of Jupiter's moons. It is about the same size as the planet Mercury. Ganymede has a mass of 1.66×10^{23} kg, an orbital radius of 1.07×10^9 m and an orbital period of 6.18×10^5 s.

a Use this information to determine the mass of Jupiter.
b Calculate the orbital speed of Ganymede.
c Calculate the gravitational force that Ganymede exerts on Jupiter.
d What is the size of the gravitational force that Jupiter exerts on Ganymede?

Physics file

One of the smallest satellites was launched in 2003 by the Canadian Space Agency. It is called the MOST (Microvariability and Oscillations of Stars) and is the size of a small suitcase. It has a telescope with a 15 cm diameter mirror that is collecting images from nearby stars that are similar to our Sun. Canadian scientists have called it the Humble Space Telescope.

Solution

a To calculate the mass of Jupiter, use:

$T = 6.18 \times 10^5$ s

$G = 6.67 \times 10^{-11}$ N m^2 kg^{-2}

$r_G = 1.07 \times 10^9$ m

$$F_c = F_g$$

$$\frac{m4\pi^2 r}{T^2} = \frac{GMm}{r^2}$$

$$M = \frac{4\pi^2 r^3}{GT^2}$$

$$= \frac{4\pi^2 (1.07 \times 10^9)^3}{(6.67 \times 10^{-11})(6.18 \times 10^5)^2}$$

$$= 1.90 \times 10^{27} \text{ kg}$$

b The orbital speed of Ganymede is:

$T = 6.18 \times 10^5$ s

$r_G = 1.07 \times 10^9$ m

$$v = \frac{2\pi r}{T}$$

$$= \frac{2\pi(1.07 \times 10^9)}{6.18 \times 10^5}$$

$$= 1.09 \times 10^4 \text{ m s}^{-1}$$

i.e. about 11 km s^{-1}

c The gravitational force that Ganymede exerts on Jupiter is:

$m_G = 1.66 \times 10^{23}$ kg

$G = 6.67 \times 10^{-11}$ N m^2 kg^{-2}

$r_G = 1.07 \times 10^9$ m

$M_J = 1.90 \times 10^{27}$ kg

$$F_g = \frac{GMm}{r^2}$$

$$F_g = \frac{(6.67 \times 10^{-11})(1.90 \times 10^{27})(1.66 \times 10^{23})}{(1.07 \times 10^9)^2}$$

$$F_g = 1.84 \times 10^{22} \text{ N}$$

$$F_{G \text{ on } J} = 1.84 \times 10^{22} \text{ N}$$

d The gravitational force that Jupiter exerts on Ganymede will also equal 1.84×10^{22} N.

Physics in action — Kepler's laws

When Isaac Newton developed his law of universal gravitation, he was building on work previously done by Nicolaus Copernicus, Johannes Kepler and Galileo Galilei. Copernicus had proposed a Sun-centred (heliocentric) solar system, Galileo had developed laws relating to motion near the Earth's surface, and Kepler had devised rules concerned with the motion of the planets.

Kepler, a German astronomer, published his three laws on the motion of planets in 1609, about 80 years before Newton's law of universal gravitation was published. These laws are as follows.

1. The planets move in elliptical orbits with the Sun at one focus.
2. The line connecting a planet to the Sun sweeps out equal areas in equal intervals of time.
3. For every planet, the ratio of the cube of the average orbital radius to the square of the period of revolution is the same,

i.e. $\frac{r^3}{T^2} = \text{constant}$.

The most significant consequence of Kepler's laws was that planets were no longer considered to move in perfect circles at constant speeds. His first two laws proposed that planets moved in elliptical paths, and that the closer they were to the Sun, the faster they moved. It took Kepler many months of laborious calculations to arrive at his third law. Newton used Kepler's laws to justify the inverse square relationship. In fact, Kepler's third law can be deduced, for circular orbits, from Newton's law of universal gravitation:

$$F_g = F_c$$

$$\frac{GMm}{r^2} = \frac{4\pi^2 rm}{T^2}$$

$$\frac{r^3}{T^2} = \frac{GM}{4\pi^2}$$

That is, for a given central body of mass M, the ratio $\frac{r^3}{T^2}$ is constant and equal to the collection of constants $\frac{GM}{4\pi^2}$ for all of its satellites. So, for example, if you know the orbital radius, r, and period, T, of one of the moons of Saturn, you could calculate $\frac{r^3}{T^2}$ and use this as a constant value for all of Saturn's moons. If you knew the period, T', of a different satellite of Saturn, it would then be straightforward to calculate its orbital radius, r'.

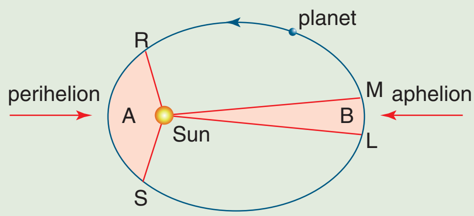


Figure 2.17
Planets orbit in elliptical paths with the Sun at one focus. Their speed varies continuously, and they are fastest when closest to the Sun. A line joining a planet to the Sun will sweep out equal areas in equal times, e.g. the time it takes to move from R to S is equal to the time it takes to move from L to M, and so area A is the same as area B.

Physics in action — What is a planet?

Pluto—the tribe has spoken

In 2002, another world in orbit around our Sun was discovered. Quaoar (Kwah-o-ar) is further out than Pluto, takes 288 years to orbit and is about 1250 km in diameter. Then in 2005, Xena, a similar object but larger than Pluto, was also discovered. Both Xena and Quaoar are ice dwarfs that inhabit the Kuiper belt, a region way beyond the orbit of Neptune. Astronomers have found about 40 of these dwarfs in this region and will probably find many more.

This led to a controversial debate among astronomers about whether all these distant ice dwarfs, including Pluto, should be called planets. In 2006, at a meeting of the International Astronomical Union, it was decided that a planet should be large enough to become spherical under its own gravity and should be large enough to dominate the region. Pluto failed this test because its moon Charon is almost as large as Pluto. Xena, Quaoar, Charon and Pluto are now officially classified as dwarf planets.

table 2.4 Data for the Sun, its eight planets and Earth's moon

Body	Mass (kg)	Radius (m)	Period of rotation	Mean orbital radius (m)	Period of orbit	Av. orbital speed (km s^{-1})
Sun	1.98×10^{30}	6.95×10^{8}	24.8 days	NA	NA	NA
Mercury	3.28×10^{23}	2.57×10^{6}	58.4 days	5.79×10^{10}	88 days	47.8
Venus	4.83×10^{24}	6.31×10^{6}	243 days	1.08×10^{11}	224.5 days	35.0
Earth	5.98×10^{24}	6.38×10^{6}	23 h 56 min	1.49×10^{11}	365.25 days	29.8
Mars	6.37×10^{23}	3.43×10^{6}	24.6 h	2.28×10^{11}	688 days	24.2
Jupiter	1.90×10^{27}	7.18×10^{7}	9.8 h	7.78×10^{11}	11.9 years	13.1
Saturn	5.67×10^{26}	6.03×10^{7}	10 h	1.43×10^{12}	29.5 years	9.7
Uranus	8.80×10^{25}	2.67×10^{7}	10.8 h	2.87×10^{12}	84.3 years	6.8
Neptune	1.03×10^{26}	2.48×10^{7}	15.8 h	4.50×10^{12}	164.8 years	6.5
Moon	7.34×10^{22}	1.74×10^{6}	27.3 days	3.8×10^{8}	27.3 days	1.0

Physics in action — Case studies: Three satellites

Geostationary Meteorological Satellite (MT SAT-1R)

The Japanese MT SAT-1R satellite was launched in February 2005 and orbits at 35 800 km directly over the Equator. Its closest point to the Earth, or perigee, is 35 776 km. Its furthest point from the Earth, known as the apogee, is at 35 798 km. MT SAT-1R orbits at a longitude of 140°E, so it is just to the north of Cape York and ideally located for use by Australia's weather forecasters. Signals from MT SAT-1R are transmitted 2-hourly and are received by a satellite dish on the roof of the Bureau of Meteorology head office in Melbourne. Infrared images show the temperature variations in the atmosphere and are invaluable in weather forecasting. MT SAT-1R is box-like and measures about 2.6 m each side. It has a mass of 1250 kg and is powered by solar panels that when deployed take its overall length to over 30 m.

Hubble Space Telescope (HST)

This cooperative venture between NASA and the European Space Agency (ESA) was launched by the crew of the Space Shuttle *Discovery* on 25 April 1990. HST is a permanent unoccupied space-based observatory with a 2.4 m diameter reflecting telescope, spectrographs and a faint-object camera. It orbits above the Earth's atmosphere and so produces images of distant stars and galaxies far clearer than those from ground-based observatories. HST is in a low-Earth orbit inclined at 28° to the Equator. Its forecast life span was about 15 years, but a successful service and repair mission carried out in 2009 is expected to extend its life by another 5 years.

National Oceanic and Atmospheric Administration Satellite (NOAA-18)

The US-owned and operated NOAA satellites are located in low-altitude polar orbits. This means that they pass over the poles of the Earth as they orbit. NOAA-18 was launched in May 2005 and orbits at an inclination of 99° to the Equator. Its low altitude means that it captures high-resolution pictures of small bands of the Earth. The data is used in local weather forecasting as well as provide enormous amounts of information for monitoring global warming and climate change.

Most of the satellites that are in orbit are used for military surveillance. These are top-secret projects and not much is known about their orbital properties. They use high-powered telescopes and cameras to observe the Earth and transmit pictures to their home country.

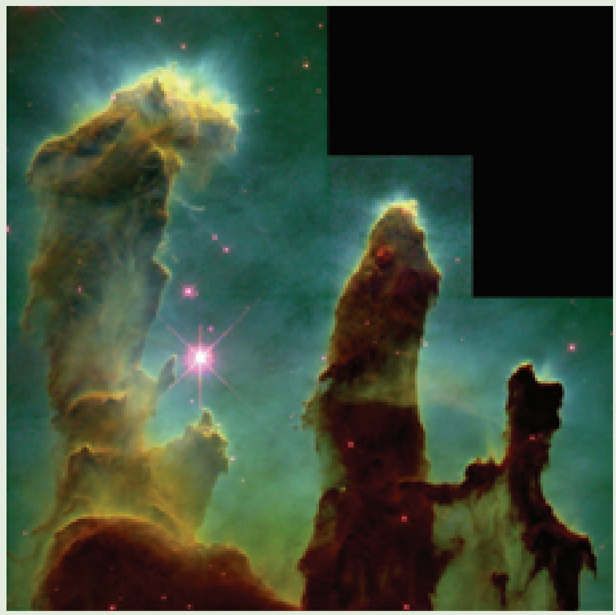


Figure 2.18
This picture was received from the Hubble Space Telescope in 1995. It shows vast columns of cold gas and dust, 9.6×10^{12} km long, in the constellation Serpens, just 7000 light-years away.

table 2.5 Data for three satellites that orbit the Earth

Satellite	Orbit	Inclination	Perigee (km)	Apogee (km)	Period
MT SAT-1R	Equatorial	0°	35 776	35 798	1 day
HST	Inclined	28°	591	599	96.6 min
NOAA-18	Polar	99°	846	866	102 min

2.2 SUMMARY Satellite motion

- A satellite is an object that is in a stable orbit around a more massive object.
- The Solar System contains many natural satellites. The planets are natural satellites of the Sun and the moons are natural satellites of the planets.
- Many artificial satellites have been placed in orbit around the Earth. Some artificial satellites have been placed in orbit around other planets.
- The only force acting on a satellite is the gravitational attraction between it and the central body.
- Satellites are in continual free-fall. They move with a centripetal acceleration that is equal to the gravitational field strength at the location of their orbit.
- The speed of a satellite is given by:

$$v = \frac{2\pi r}{T}$$

- The acceleration of a satellite in a circular orbit is given by:

$$\boldsymbol{a}_c = \frac{v^2}{r} = \frac{4\pi^2 r}{T^2} = \frac{GM}{r^2} = \boldsymbol{g}$$

- The gravitational force acting on a satellite in a circular orbit is given by:

$$\boldsymbol{F}_c = m\boldsymbol{a}_c = \frac{mv^2}{r} = \frac{m4\pi^2 r}{T^2} = \frac{GMm}{r^2} = m\boldsymbol{g} = \boldsymbol{F}_g$$

- For any central body of mass M:

$\frac{r^3}{T^2} = \frac{GM}{4\pi^2} =$ constant for all satellites of this central body. So knowing another satellite's orbital radius, r, enables its period, T, to be determined.

2.2 Questions

1 Which of the following statements is correct?
A satellite in a stable circular orbit 100 km above the Earth will move with:
- A an acceleration of 9.8 m s^{-2}
- B a constant velocity
- C zero acceleration
- D an acceleration of less than 9.8 m s^{-2}.

2 Explain why the gravitational field of the Earth does no work on a satellite in a stable circular orbit.

The following information applies to questions 3 and 4.
The gravitational field strength at the location where the Optus D1 satellite is in stable orbit around the Earth is equal to 0.220 N kg^{-1}. The mass of this satellite is 2.30×10^3 kg.

3 Using only the information given, calculate the net force acting on this satellite as it orbits.

4 Identify the source of this net force.

5 The planet Neptune has a mass of 1.02×10^{26} kg. One of its moons, Triton, has a mass of 2.14×10^{22} kg and an orbital radius equal to 3.55×10^8 m. For Triton, calculate its:
- a orbital acceleration
- b orbital speed
- c orbital period (in days).

The following information applies to questions 6 and 7.
One of Saturn's moons, Titan, has a mass of 1.35×10^{23} kg and an orbital radius of 1.22×10^9 m. The orbital period of Titan is 1.38×10^6 s.

6 Calculate the:
- a orbital speed of Titan (in km s^{-1})
- b orbital acceleration of Titan.

7 Using this data, calculate the mass of Saturn.

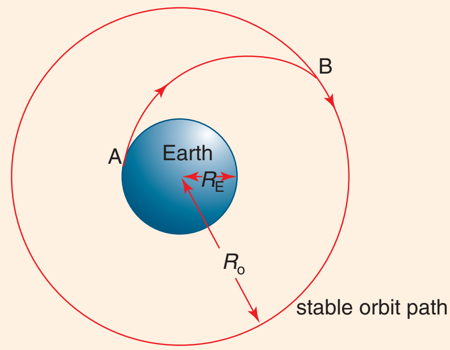
8 A satellite is in a geosynchronous orbit around the Earth if its period of rotation is the same as that of the Earth, i.e. 24 h. Such a satellite is called a geostationary satellite. Venus has a mass of 4.87×10^{24} kg and a radius of 6.05×10^6 m. The length of a day on Venus is 2.10×10^7 s. For a satellite to be in a synchronous orbit around Venus, calculate:
- a the orbital radius of the satellite
- b its orbital speed
- c its orbital acceleration.

9 The data for two of Saturn's moons, Atlas and Helene, is as follows. The orbit of Helene is about twice as far from the centre of Saturn as that of Atlas.

	Orbital radius (m)	Orbital period (days)
Atlas	1.37×10^8	0.602
Helene	3.77×10^8	2.75

- a Calculate the value of these ratios:
  - i $\frac{\text{orbital speed of Atlas}}{\text{orbital speed of Helene}}$
  - ii $\frac{\text{acceleration of Atlas}}{\text{acceleration of Helene}}$
- b The largest of Saturn's moons is Titan. It has an orbital radius of 1.20×10^9 m. Use Kepler's third law to show that the orbital period of Titan is 15.6 days.

10 The Space Shuttle is launched into orbit from a point A on the Equator, as shown. The shuttle then enters a stable circular orbit of radius r_o at point B. The radius of the Earth is 6.4×10^6 m. The ratio of the gravitational field strength at A to that at B is equal to 1.2. Calculate the distance r_o.



2.3 Torque

So far in your study of physics you have generally analysed motion in a straight line: even projectile and circular motion has been presented in terms of linear quantities. However, many real-life situations involve bodies that *rotate*—closing a door, using a spanner or screwdriver, turning the volume knob on an amplifier, opening a bottle of soft-drink, and so on. In these situations, a linear force acts to provide a *turning effect*, or more precisely, a *torque*. Newton's laws use the concept of a force to help understand the motion of a body in a straight line. The concept of torque is used in exactly the same way to explain a change in the *rotational motion* of a body or system of bodies. A torque is required in any situation in which a body is caused to rotate.

Figure 2.19
Some situations in which a torque is acting. In each case, a force ***F*** is applied at a distance *r* from a pivot point, resulting in a rotational or turning effect.

The amount of torque, τ, created by a force will depend on three factors:

- the magnitude of the force, *F*. A larger force will result in a larger torque
- the perpendicular distance from the line of action of the force to the pivot point ($r_\perp$). The greater the perpendicular distance, the greater the torque
- the angle, θ, between the lever arm and the force, which affects the size of the perpendicular distance, $r_\perp$.

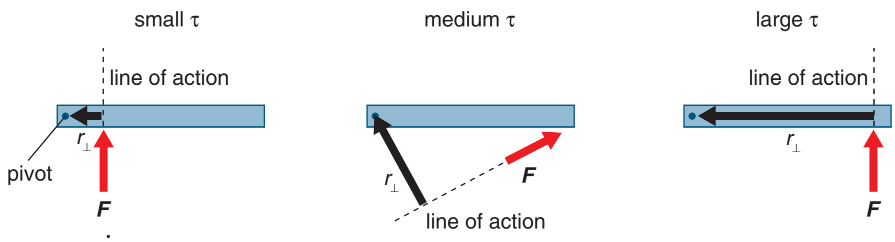


Figure 2.20

The size of a torque depends on the perpendicular distance ($r_\perp$) from the line of action of the force to the axis rotation. This is affected by the angle, θ, that the force makes with the lever arm.

When analysing a rotating system, the axis of rotation is an important reference point. A door, for example, moves in a circular arc around its hinges. The line of the hinges is the axis of rotation. A force applied at the hinge itself will not create a turning effect as $r_\perp$ will be zero. The maximum effect will be achieved by applying a force to the door as far from the hinge as possible in order to maximise $r_\perp$. Similarly, when a spanner is used to tighten a nut, the centre of the nut is the axis of rotation, and a spanner with a longer handle will provide a greater torque on the nut.

The turning effect of a force also depends on the *direction* in which it is applied. In closing a door, for example, the maximum effect is achieved if the force you apply is at 90° to the door surface. This results in $r_\perp$ becoming the largest value possible. If this angle is reduced, there is a smaller perpendicular component of the radius from the line of action of the force to the door hinge and so a smaller torque is produced. If the force is directed along the line of the door (i.e. 0°, directly towards the hinges), $r_\perp$ will be zero and the door will not rotate, no matter how great a force you apply.

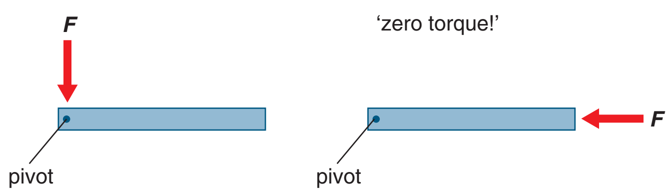


Figure 2.21

No torque is created in these situations. The forces do not generate any rotational or turning effect.

The **_TORQUE_** (τ) acting on a body is given by the product of the applied force acting on the lever arm, ***F***, and the perpendicular distance from the line of action of the force to the axis of rotation, $r_\perp$:

$$\tau = Fr_\perp$$

where τ (tau) is the torque (N m), ***F*** is the force applied to the lever arm (N) and $r_\perp$ is the perpendicular distance from the line of action of the force to the pivot point or axis of rotation (m). Torque is a vector quantity and must be assigned a direction, usually clockwise or anticlockwise.

A torque can result in a rotation that can be given as either clockwise or anticlockwise. An anticlockwise rotation can be considered to be positive and a clockwise rotation can be taken as negative. This approach is useful when a number of torques are acting on a body and the net effect has to be found.

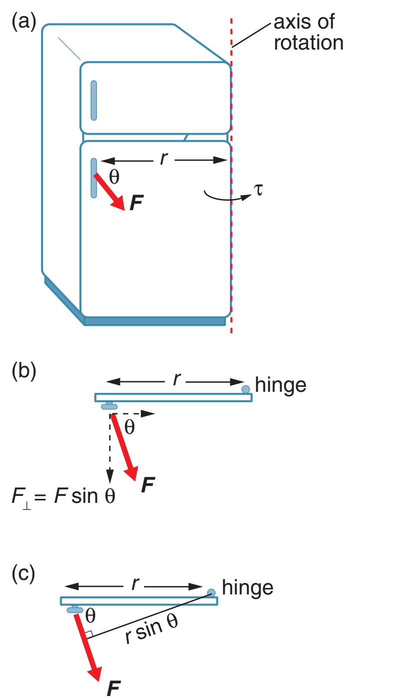


Figure 2.22

(a) A torque applied to a door acts around the axis of rotation (the hinges). (b) This torque is supplied by a force with a component $F_\perp = F \sin\theta$ acting perpendicular to the door at a distance r from the axis of rotation, i.e. $\tau = F_\perp r$. (c) The torque is also the product of the applied force ***F*** and perpendicular distance from the force to the hinges, $r_\perp = r \sin\theta$, i.e. $\tau = Fr_\perp$.

Physics file

Although the unit for torque (N m) appears to be the same as that for work (1 J = 1 N m), it is important to realise that they are very different units. Work ($W = F \times x$) involves a force and the distance over which the force is acting. With work, the force has a component in the direction of the motion. Torque ($\tau = Fr_{\perp}$) is the product of a force and the perpendicular distance at which that force is acting from a given point. In the case of torque the force acts in the direction of the rotation.

✓ Worked Example 2.3A

A woman whose car has a flat tyre has two wheel-nut spanners in the boot of her car. One wheel spanner is 15.0 cm long and the other is 75.0 cm long.

a In order to undo the wheel nuts with a minimum amount of effort, which wheel spanner should the woman select?

b If the maximum force that the woman can apply is 45.0 N, determine the maximum torque that can be delivered to a wheel nut.

Solution

a The woman should choose the longer wheel spanner. The longer lever arm means that the force is applied at a greater distance from the axis of rotation. Since $\tau = Fr_{\perp}$, the longer lever will deliver a larger torque to the wheel nut. Using a pipe to make the lever arm even longer gives even more torque!

b Maximum torque will be obtained by applying the 45.0 N force perpendicular to the lever arm at the maximum distance of 75 cm from the wheel nut. Thus:

$F = 45.0\,\text{N}$

$r_{\perp} = 75.0 \times 10^{-2}\,\text{m}$

$$\tau = Fr_{\perp}$$
$$= (45.0)(75.0 \times 10^{-2})$$
$$= 33.8\,\text{N m}$$

Physics in action — Centre of mass and stability

Think about an athlete running in a 100 m sprint. In simple terms, the athlete runs in a straight line along the track, and her displacement and velocity at any time can be calculated using the principles discussed in section 1.2. In reality, however, the motion of the various parts of the athlete's body will differ significantly during the run. Her arms and legs move in a complex manner that is not easy to analyse.

The analysis of the motion of complicated systems such as a sprinter or high-jumper can be simplified to the motion of a single point. The mass of the sprinter can be considered to be 'concentrated' into a point which has travelled in a straight line. This single point is called the *centre of mass*. A most important property of the centre of mass is that it will follow a path that is exactly the same as the path of a point particle of the same mass if it were subjected to the same net force.

If a body is uniform in one dimension only (e.g. a straight piece of wire), its centre of mass will lie exactly at the centre. In two dimensions, the centre of mass will be the point which is central for both dimensions. It is even possible for the centre of mass to lie outside the body, as with a doughnut (the centre of mass is in the hole). A person's centre of mass is typically just below the chest, but it will vary with the positions of the arms and legs.

A concept that is closely related to centre of mass is *centre of gravity*. Instead of being a point particle whose motion equates to the whole extended body or system, the centre of gravity is the position from which the entire *weight* of the body or system is considered to act. As a consequence of this, the centre of gravity is the position at which the body will balance. For all practical purposes, the centre of gravity is exactly at the centre of mass. It is only when a body is so large that it is in a non-uniform gravitational field that the centre of gravity no longer coincides with the centre of mass.

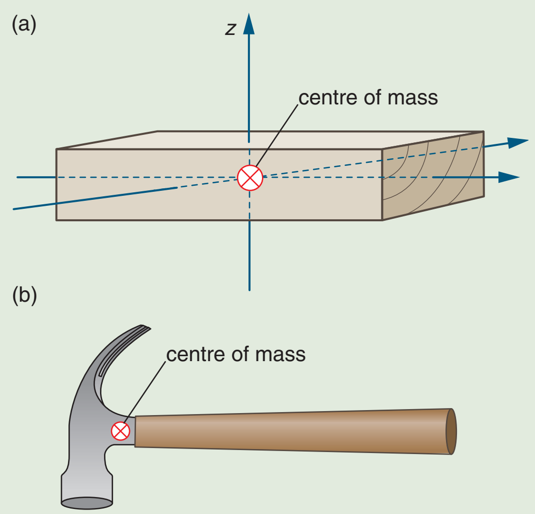


Figure 2.23
The centre of mass for (a) a uniform body in three dimensions and (b) a hammer

In designing structures, engineers and architects want to ensure that balance and stability are maintained. This will depend on the relative positions of the centre of gravity and the base or point of support. When a vertical line downwards from the centre of gravity passes through the base of support, the object is stable. The vertical line from the centre of gravity represents the direction of the force of gravity on the object. In Figure 2.24a the weight of the car passes through the car's support base. The torque acting on the car therefore does not cause the car to tip over. In the case of the truck, however, the weight is directed outside the point (base) of support, so the torque acts to tip the truck over.

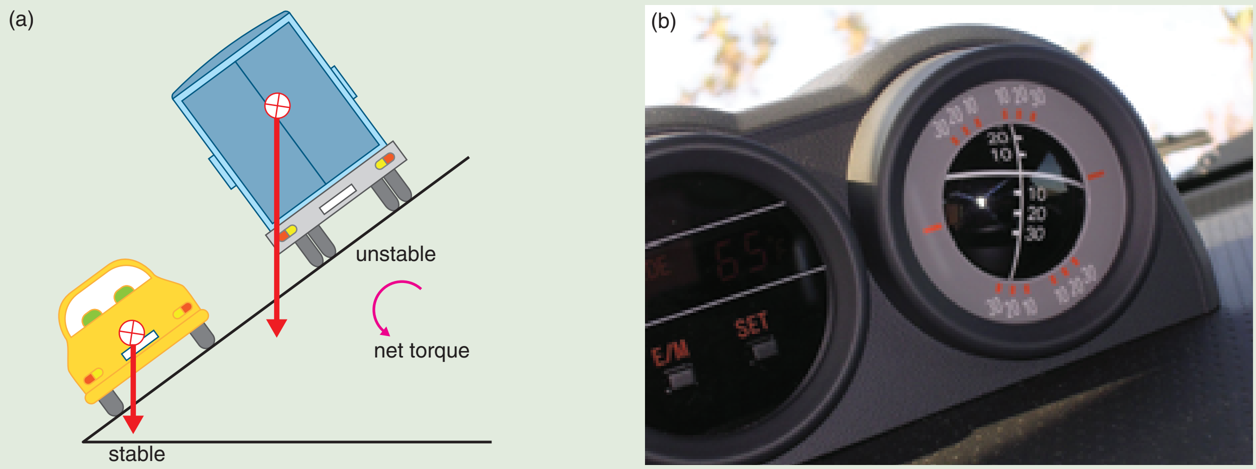


Figure 2.24
(a) The car on the incline is in stable equilibrium, while the heavily laden truck on the same incline could topple. The weight vector is outside the lower point of support for the truck, so there is no reaction force from the road to the higher wheel. (b) Modern four-wheel drives and tractors have inclinometers to warn the driver if the vehicle is in danger of tipping.

The stability of an object or structure can be increased in a number of ways. If the centre of gravity is lowered or the width of the support base is increased, the angle from the centre of gravity to the edge of the base is increased. As a result, the object has to be tipped further to make the force of gravity act outside the support base. Racing cars have a very low centre of gravity to increase their stability when cornering at high speed. In a similar way, training wheels on a child's bicycle widen the support base, making it harder to tip the bicycle sideways.

Figure 2.25
The owner of this cart did not have a good understanding of stability and balanced torques!

2.3 SUMMARY Torque

- A force that acts to cause a rotation is said to provide a turning effect or torque, τ.
- The torque acting on a body is given by the product of the applied force on the lever arm, F, and the perpendicular distance from the line of action of the force to the axis of rotation, $r_{\perp}$:
$$\tau = Fr_{\perp}$$
- A force acting directly towards or away from the pivot point produces no torque.

2.3 Questions

Use g = 9.80 N kg^{-1} when answering these questions.

1 In the following situations, a torque is acting. In each case, identify the axis of rotation or pivot point about which the torque acts and also estimate the length of the lever arm.
  a A garden tap is turned on.
  b A wheelbarrow is lifted by the handles.
  c An object is picked up with a pair of tweezers.
  d A screwdriver is used to lever open a tin of paint.

2 Use the concept of torque to explain the following.
  a It is easier to open a heavy door by pushing it at the handle than in the middle of the door.
  b It is possible to move very heavy rocks in the garden by using a long crowbar.

3 Halina and Stefanie, each of mass 40.0 kg, are sitting at opposite ends of a playground see-saw. The see-saw is stationary in the horizontal position and is 4.50 m long. Stefanie decides to jump off. Calculate the size of the unbalanced torque that now acts on the see-saw.

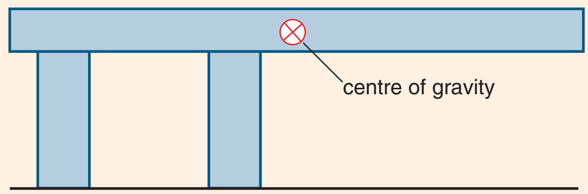
4 A wheelbarrow measures 1.60 m from the tips of the handles to the wheel axle. The contents of the wheelbarrow produce a clockwise torque of 400.0 N m about the wheel axle. What is the smallest force that must be applied at the handles to lift the wheelbarrow so that it is ready for moving? Assume the force acts at 90.0° to the lever arm.

5 A crane with a horizontal lever arm is lifting a concrete wall of mass 2.50 tonnes. The load is 2.00 m from the axis of rotation.
  a Calculate the torque created by this load.
  b What stops the crane from toppling over as a result of this torque?

6 Nikki is investigating torque using a metre rule and a 1.00 kg mass. She uses a rubber band to attach the mass to the ruler. Nikki first holds the ruler at one end so that it is horizontal, with the mass at the 50.0 cm mark.
  a What is the size of the torque that is acting?
  b She now moves the mass so that it is right at the far end of the ruler. How much torque is acting now?
  c Finally, she lifts the ruler so that it makes an angle of 60.0° to the horizontal. What is the size of the torque now?

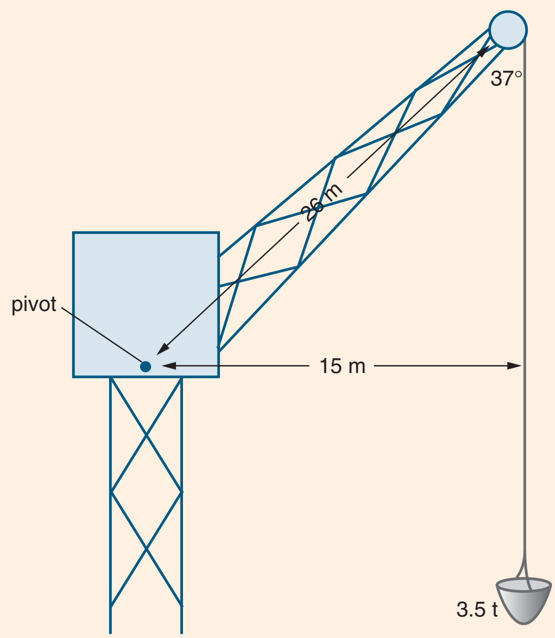
7 Tight-rope walkers sometimes carry a long balancing pole with ends that extend below the level of the rope. The poles often carry weights at their ends. Consider the torques that act here and explain how they help the performer to remain in stable equilibrium were they to overbalance.

8 Christopher likes constructing things, but he does not have a good understanding of stability. He draws a design for a garden bench that consists of a long beam resting on top of two supports, as shown in the diagram. The centre of gravity of the beam is indicated. Christopher does not intend to use any nails, screws, bolts, ropes or adhesives in his bench. Explain whether the bench will work successfully and what he should do to improve the design.



9 a When you carry a heavy bag in your right hand, you automatically raise your left arm. Why is this?
  b Estimate the magnitude of the torque that a 14.0 kg suitcase held in your right hand would exert on your body, if your spine is the axis of rotation.

10 A crane is being used to lift a skip of concrete with a total mass of 3.50 tonnes. The lever arm of the crane is 25.0 m long and makes an angle of 37.0° with the vertical as shown in the diagram. Ignore the mass of the cable when answering these questions.
  a What is the total weight of the skip?
  b The skip is lifted so that it is near the top of the crane. How does the torque created by the skip about the pivot change as the skip is lifted to this height?
  c Calculate the magnitude of the torque about the pivot that the skip exerts on the crane when the skip is at the highest point.



2.4 Equilibrium

When all the forces acting on an object add up to a zero net force, the body is said to be in *equilibrium* (or more accurately, *translational equilibrium*). It does not matter whether the object is as small as a jelly bean or as large as a skyscraper. The object might be moving (possibly even rotating), but there will be no translational acceleration if the forces acting on it are balanced. This is the situation described by Isaac Newton in his first law of motion: if the forces acting on a body are balanced, a stationary object will remain stationary and a moving object will keep moving with a constant velocity.

For example, a book lying on a table has two forces acting upon it: the force due to gravity downwards and a normal force upwards. The net force is zero, so the book remains at rest. Another example is an aircraft travelling in a straight line. Four forces act on the aircraft: the lift upwards (from the wings), the force due to gravity on the aircraft downwards, drag backwards (from air resistance) and the thrust forwards (provided by the jet engines). As long as these forces add to a zero net force, the aircraft will move with a constant velocity; that is, it will be in a state of translational equilibrium. If any one of these forces changed, the aircraft would not be in translational equilibrium and would accelerate in the direction of the net force.

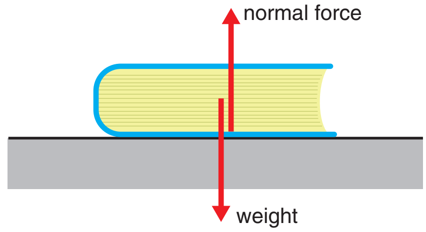


Figure 2.26

A book lying on a table has two forces acting upon it: the force due to gravity downwards and a normal force upwards. The net force is zero, so the book remains at rest.

> A body is said to be in ***TRANSLATIONAL EQUILIBRIUM*** when the sum of the forces acting on the body is zero, i.e. $\Sigma \boldsymbol{F} = 0$.

Since force is a vector quantity, the separate *components* of the net force on a body must also be zero for it to be in equilibrium. In two-dimensional situations, this means that $\Sigma F_x = 0$ and $\Sigma F_y = 0$ as well. This fact can be most useful in solving some problems. For example, when we consider a moving bicycle, the forces in the *y*-direction, its weight and the normal force supplied by the road, can be ignored as they satisfy the condition $\Sigma F_y = 0$.

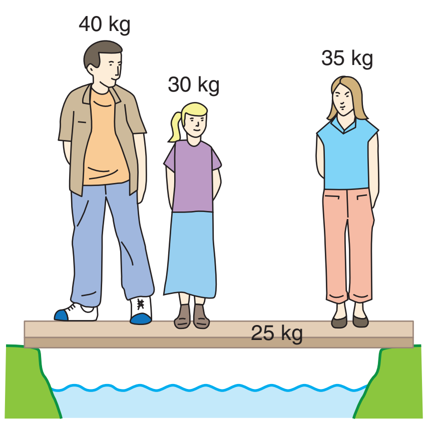
✓ Worked Example 2.4A

Three children are standing on a plank that is bridging a small stream. The plank is supported at each end by the ground. The plank has a mass of 25 kg and the children have masses of 40, 30 and 35 kg.

Use $\boldsymbol{g} = -9.80$ N kg^{-1} when answering this question.

a Draw a free-body diagram showing all the forces that are acting on the plank.

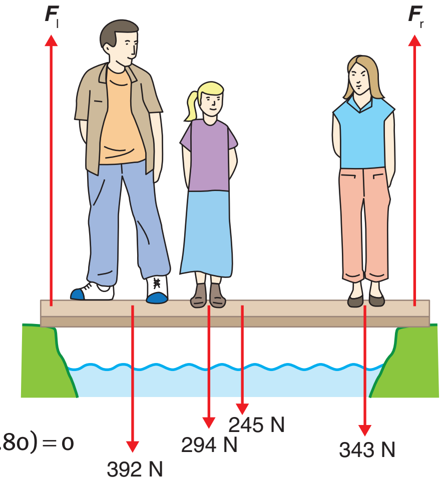
b At the left-hand end of the plank, the ground exerts an upward force of 700 N on the plank. What is the magnitude of the force that the ground exerts on the plank at the right-hand end?



Solution

a There are six forces acting on the plank. Upward forces from the ground act at each end. Downward forces from each of the children, equal to their weight, are acting. The weight of the plank itself, acting at the centre of mass of the plank, should not be forgotten. These forces are shown in the diagram. It is important that these force vectors are drawn so that they are in proportion, in the correct location, and in with the tails of the vectors in contact with the plank. Your free-body diagram should not include the children, supports or water.

b The forces acting on the plank are in equilibrium, i.e. $\Sigma \boldsymbol{F} = 0$. This means that the upward forces on the plank are being balanced by the downward forces.



$\updownarrow^{+}_{-}$ $\quad \Sigma \boldsymbol{F} = 0$

$g = -9.80\ \text{m s}^{-2}$ $\quad \therefore \boldsymbol{F}_l + \boldsymbol{F}_r + \boldsymbol{F}_{ch_1} + \boldsymbol{F}_{ch_2} + \boldsymbol{F}_{ch_3} + \boldsymbol{F}_{plank} = 0$

$(700) + \boldsymbol{F}_r + (40)(-9.80) + (30)(-9.80) + (35)(-9.80) + (25)(-9.80) = 0$

$(700) + \boldsymbol{F}_r + (-392) + (-294) + (-343) + (-245) = 0$

$\boldsymbol{F}_r + (-574) = 0$

$\boldsymbol{F}_r = +574\ \text{N}$

$= 574\ \text{N upwards}$

The concept of forces in equilibrium is an essential part of *statics*, the branch of physics devoted to the study of objects and structures in equilibrium. Combined with an understanding of the properties of materials, statics is most important to architects and engineers when designing safe and functional buildings, machines, bridges and other structures.

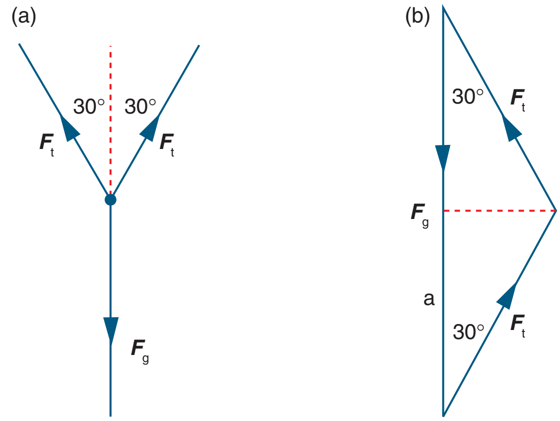
Within a structure, a rigid body such as a beam or column may have many forces acting on it in many different directions. However, the force of gravity will always act vertically downwards as though applied to the centre of gravity of the object. The centre of gravity of any uniform rigid body that is symmetrical in three dimensions is always at the centre of the body. Structures may also include flexible components such as supporting cables, chains or ropes. Because these elements are not rigid, forces must act *along* them. If this were not the case, a supporting rope (for example) would bend or buckle and would no longer act as a support.

✓ Worked Example 2.4B

A rigid advertising banner is to be hung by two guy ropes. The banner has a mass of 45 kg and the ropes must be at an angle of 30° to the vertical, as shown. If the mass of the ropes is ignored, determine the tension in each rope required to support the banner.

Solution

The banner is stationary and so the forces acting on it are in equilibrium, i.e. $\Sigma F = 0$. The forces acting on the banner are the force due to gravity, F_g, of the banner (which acts at the centre of gravity) and the force supplied by the tension in each rope. A force diagram (a) can be used to depict the direction of the forces. The tension, F_t, in each rope can be found by constructing a vector diagram representing the forces adding to zero, as in (b).



Using the right-angle triangle in (b):

$a = \frac{F_g}{2}$ $\qquad h = \frac{a}{\cos 30°}$

$h = F_t$ $\qquad F_t = \frac{F_g}{2\cos 30°}$

$m = 45\text{ kg}$ $\qquad = \frac{(45)(9.80)}{2\cos 30°}$

$g = 9.80\text{ N kg}^{-1}$ $\qquad = 2.55 \times 10^2$ N along the rope

The tensile force is 250 N. Note that the gravitational field strength's negative direction is not needed as the direction of the tension force is always along the rope.

Structures in translational and rotational equilibrium

As we saw in the previous section, the forces acting on an object can be equal and opposite, yet the object is not in equilibrium. For example, a mechanic might use a T-shaped spanner to undo a wheel nut. If the mechanic applies equal and opposite forces to the arms of the spanner, the spanner does not remain at rest or move with a constant velocity. It *rotates*.

This shows that a system for which the sum of all the forces is zero (i.e. $\Sigma F = 0$) may not always be in equilibrium. Another example is a racing cyclist riding a bike with clip-in pedals. The cyclist's feet provide a pair of opposite forces that produce a torque about the axle of the pedals. Importantly, both of the forces deliver a torque that acts in the same sense (i.e. they are either both clockwise or both anticlockwise), creating a net torque so that the pedals rotate. Forces that act in this way are said to be a *couple*. Even though the sum of the forces is zero, a body that is subject to a couple is not in equilibrium.

Practical activity

16 Force and equilibrium

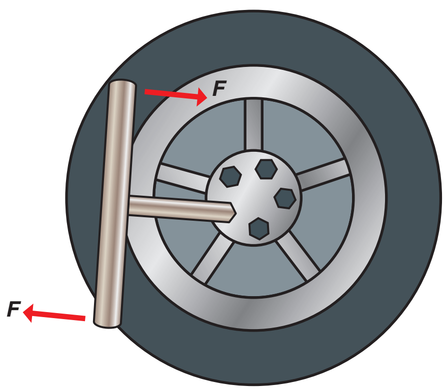


Figure 2.27
Equal and opposite forces are acting on the wheel spanner, but it is not in equilibrium.

Rotational equilibrium

For a structure to be completely at rest and stable, it is not enough that it is simply in translational equilibrium. The structure or system must also be in *rotational equilibrium*. Simply stated, this means that the sum of all the torques acting about a point must be zero, or $\Sigma\tau = 0$. This means that the net clockwise torque must be equal in magnitude to the net anticlockwise torque; that is, $\Sigma\tau_{\text{clockwise}} = \Sigma\tau_{\text{anticlockwise}}$.

For a body or system to be in ***ROTATIONAL EQUILIBRIUM***, the sum of all the torques acting about a point must be zero:

$$\Sigma\tau = 0$$

or

$$\Sigma\tau_{\text{clockwise}} = \Sigma\tau_{\text{anticlockwise}}$$

Figure 2.28

The cyclist may exert equal and opposite forces on the pedals, but this system is not in static equilibrium. A force couple is produced if the forces produced by each foot are applied to different points, causing a pair of torques that act in the same sense. The pedals rotate as a result of this couple.

Practical activity

15 Seesaws

Static equilibrium

When a body or system is not accelerating or rotating, it is in *both* translational and rotational equilibrium (i.e. $\Sigma F = 0$ and $\Sigma\tau = 0$), and so it is said to be in static equilibrium. A building should be in static equilibrium.

For a body or system to be in ***STATIC EQUILIBRIUM***, it must be in both translational and rotational equilibrium:

$$\Sigma \mathbf{F} = 0 \text{ and } \Sigma\tau = 0$$

✓ Worked Example 2.4C

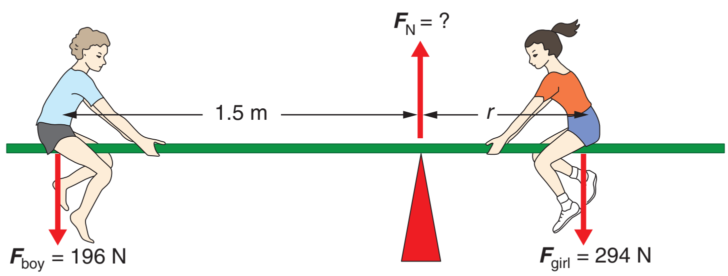
While playing in their backyard, two young children make a see-saw with a long plank. The boy sits on the see-saw 1.5 m from the pivot. The girl decides to see where she has to sit in order to balance the boy. The mass of the boy and girl are 20 kg and 30 kg, respectively. Assume that the plank's mass is negligible.

a What is the force supplied to the plank by the pivot when both children are sitting on the plank?

b Where must the girl sit in order to balance the boy?

Solution

a The plank is in equilibrium, and a force diagram reveals that there are three forces acting on the plank. The two weights are known, so the third (the upwards force on the plank supplied by the pivot) can be found.



$\updownarrow^{+}_{-}$

$g = -9.80 \text{ m s}^{-2}$

$$\Sigma \mathbf{F} = 0$$
$$\therefore \mathbf{F}_{boy} + \mathbf{F}_{girl} + \mathbf{F}_{pivot} = 0$$
$$(20)(-9.80) + (30)(-9.80) + \mathbf{F}_{pivot} = 0$$
$$(-196) + (-294) + \mathbf{F}_{pivot} = 0$$
$$\mathbf{F}_{pivot} + (-490) = 0$$
$$\mathbf{F}_{pivot} = +490 \text{ N}$$
$$= 490 \text{ N upwards}$$

b For the boy and girl to be in balance, the clockwise torque caused by the girl must equal the anticlockwise torque that the boy creates. The lever arm in each case is the distance from the pivot to the child.

$\boldsymbol{F}_{girl} = -294$ N

$r_{\perp boy} = 1.5$ m

$\boldsymbol{F}_{boy} = -196$ N

$$\Sigma\tau_{cw} = \Sigma\tau_{acw}$$

$$\tau_{boy} = +\tau_{girl}$$

$$\boldsymbol{F}_{boy} r_{\perp boy} = \boldsymbol{F}_{girl} r_{\perp girl}$$

$$r_{\perp girl} = \frac{\boldsymbol{F}_{boy} r_{\perp boy}}{\boldsymbol{F}_{girl}}$$

$$= \frac{(-196)(1.50)}{(-294)}$$

$$= 1.0 \text{ m}$$

This could also be solved by assigning a positive sign to the clockwise torque and a negative sign to the anticlockwise torque, and having the sum of the clockwise and anticlockwise torques equalling zero.

In Worked Example 2.4C, the see-saw is in equilibrium because all the forces and torques are balanced. In solving the problem, it seemed obvious to choose the pivot as the point around which the torques are determined. But because the plank is in equilibrium, *any point* could have been chosen as the reference point. For example, take the reference point to be where the girl is sitting (Figure 2.29). This will mean that $\tau_{girl} = 0$, since the lever arm distance ($r_\perp$) here will be zero.

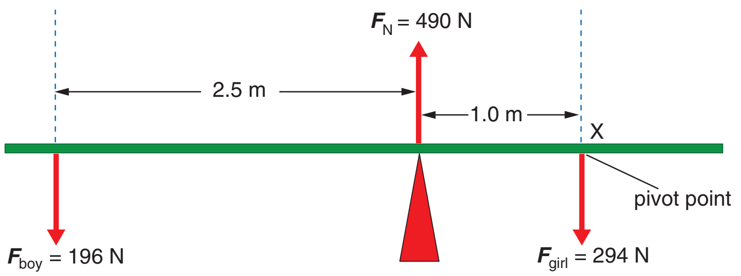


Figure 2.29

A force diagram for the see-saw problem, where the point at which the girl sits (labelled X) has been chosen as the reference point.

The boy will create an anticlockwise torque around the girl, and the normal force at the pivot for the see-saw creates a clockwise torque around the girl. This means the torque caused by the boy will be $\tau_{boy} = Fr_\perp = 196 \times 2.5 = 490$ N m anticlockwise, while the torque caused by the upwards force of the pivot on the plank will be $\tau_{pivot} = Fr_\perp = 490 \times 1.0 = 490$ N m clockwise. Clearly, these torques are equal and opposite and will balance. You can verify this further by calculating the torques around the position of the boy.

When a body or system is in ***STATIC EQUILIBRIUM***, the sum of all the torques will be zero around any point in the system.

The see-saw problem is relatively straightforward since there is only one unknown force. If there are two unknown forces, the reference point can be chosen to coincide with one of the forces. This means that it contributes no torque (since $r_\perp = 0$) and the relationship resulting from the condition $\Sigma\tau = 0$ will solve the other unknown force. Worked Example 2.4D employs this strategy.

✓ Worked Example 2.4D

While painting a tall building, a 70.0 kg painter stands 4.00 m from the left end of a 6.00 m long plank that is supported by a rope at either end. The plank has a mass of 20.0 kg. Determine the tension in each rope.

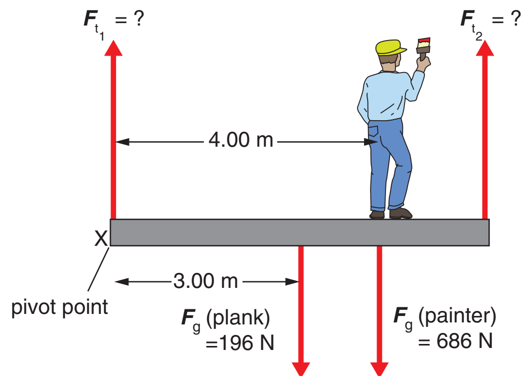
Solution

Begin by drawing a free-body diagram for the plank, which is in static equilibrium. Here, $\Sigma \boldsymbol{F} = 0$ and $\Sigma\tau = 0$.

Four forces are acting on the plank: the force due to gravity of the plank, the weight of the painter pushing down on the plank, and the tension in each of the two ropes. The force due to gravity on the plank acts from the centre of mass of the plank.

However, the equation generated from the translational equilibrium condition alone cannot be solved because there are two unknowns ($\boldsymbol{F}_{t_1}$ and $\boldsymbol{F}_{t_2}$). Torques must be considered.

One of the unknown forces can be eliminated if one of the ends of the plank is made the reference point. Let us use the left-hand end of the plank (labelled X) as the reference point for the torque analysis. The plank and the painter now provide clockwise torques and the tension in the right-hand rope provides an anticlockwise torque around X.



$m_{painter} = 70.0$ kg

$r_{\perp\, painter} = 4.00$ m

$m_{plank} = 20.0$ kg

$r_{\perp\, plank} = 3.00$ m

$r_{\perp\, t_2} = 6.00$ m

$\boldsymbol{g} = 9.80$ N kg^{-1}

$$\Sigma\tau_{cw} = \Sigma\tau_{cw}$$

$$\boldsymbol{F}_{painter} r_{\perp\, painter} + \boldsymbol{F}_{plank} r_{\perp\, plank} = \boldsymbol{F}_{t_2} r_{\perp\, t_2}$$

$$\boldsymbol{F}_{t_2} = \frac{\boldsymbol{F}_{painter} r_{\perp\, painter} + \boldsymbol{F}_{plank} r_{\perp\, plank}}{r_{\perp t_2}}$$

$$= \frac{(70.0)(9.80)(4.00) + (20.0)(9.80)(3.00)}{(6.00)}$$

$$= 5.55 \times 10^2 \text{ N upwards}$$

Similarly, the tension in the left-hand rope can be obtained by shifting the reference point to any other position. (We will use the right end of the beam here.) This time, the plank and the painter provide anticlockwise torques and the rope supplies a clockwise torque.

$m_{painter} = 70.0$ kg

$r_{\perp\, painter} = 2.00$ m

$m_{plank} = 20.0$ kg

$r_{\perp\, plank} = 3.00$ m

$r_{\perp\, t_1} = 6.00$ m

$\boldsymbol{g} = 9.80$ N kg^{-1}

$$\Sigma\tau_{cw} = \Sigma\tau_{cw}$$

$$\boldsymbol{F}_{painter} r_{\perp\, painter} + \boldsymbol{F}_{plank} r_{\perp\, plank} = \boldsymbol{F}_{t_1} r_{\perp\, t_1}$$

$$\boldsymbol{F}_{t_1} = \frac{\boldsymbol{F}_{painter} r_{\perp\, painter} + \boldsymbol{F}_{plank} r_{\perp\, plank}}{r_{\perp\, t_1}}$$

$$= \frac{(70.0)(9.80)(2.00) + (20.0)(9.80)(3.00)}{(6.00)}$$

$$= 3.27 \times 10^2 \text{ N upwards}$$

To check these values: if $\Sigma \boldsymbol{F} = 0$ then the sum of the two upward forces (tensions), 555 N + 327 N = 882 N, will add to the sum of the two downward forces, (–196) + (–686) = –882 N, to equal zero, which it does.

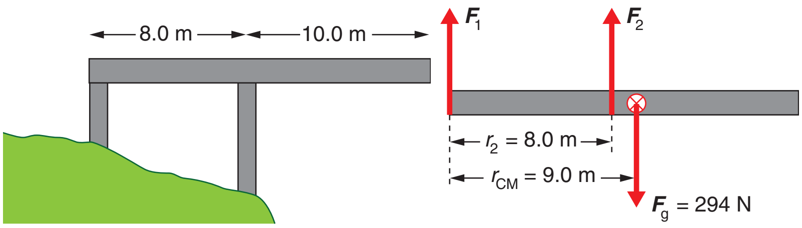
Cantilevers

A beam that extends beyond its support structure is called a *cantilever*. Cantilevers are common structural elements. For example, a cantilever bridge might be used to span a river or valley. A tower is built on each side of the river in order to support a beam projecting from each bank. Where the cantilever beams are joined at the centre of the span, there is no reduction in the force on the towers. These are the same as if the beams were not connected. All the support for the cantilever is supplied by the tower. Other structures that can involve cantilevers include shelving, roofs over the footpath outside some shops, the wing of an aircraft, and a diving board. In Perth, the Kaarta Gar-up lookout in Kings Park features a cantilevered design.

Figure 2.30
Visitors to the cantilevered Kaarta Gar-up lookout enjoy the most scenic views of the Perth city skyline.

✓ Worked Example 2.4E

A uniform cantilever beam 18.0 m long is used as a viewing platform. It extends 10.0 m beyond two supports which are 8.0 m apart. If the beam has a mass of 30 kg, determine the magnitude and direction of the force that each support must supply so that the beam is in static equilibrium.



Solution

The beam is 18.0 m long, so its centre of gravity is 9.0 m from each end. We will use the up is positive sign convention. A free-body force diagram shows the location and direction of all the forces (b). Since the beam is in rotational equilibrium, the sum of the torques will be zero. If the axis of rotation is located at the left end where $\boldsymbol{F}_1$ is acting, then there is no torque due to $\boldsymbol{F}_1$.

$m_{beam} = 30.0$ kg

$r_{\perp \text{ com beam}} = 9.00$ m

$r_{\perp F_2} = 8.00$ m

$\boldsymbol{g} = 9.80 \text{ N kg}^{-1}$

$$\Sigma\tau_{cw} = \Sigma\tau_{cw}$$

$$\boldsymbol{F}_{beam} r_{\perp \text{ com beam}} = \boldsymbol{F}_2 r_{\perp F_2}$$

$$\boldsymbol{F}_2 = \frac{\boldsymbol{F}_{beam} r_{\perp \text{ com beam}}}{r_{\perp F_2}}$$

$$= \frac{(30.0)(9.80)(9.00)}{8.00}$$

$$= 331 \text{ N upwards}$$

Rather than using the equilibrium condition for torque again, we can find $\boldsymbol{F}_1$ using the condition for translational equilibrium:

$\updownarrow^{+}_{-}$

$\boldsymbol{F}_2 = +331$ N

$\boldsymbol{F}_{beam} = -294$ N

$$\Sigma \boldsymbol{F} = 0$$

$$\therefore \boldsymbol{F}_1 + \boldsymbol{F}_2 + \boldsymbol{F}_{beam} = 0$$

$$\boldsymbol{F}_1 + (331)(-294) = 0$$

$$\boldsymbol{F}_1 + (36.8) = 0$$

$$\boldsymbol{F}_1 = -36.8 \text{ N}$$

$$= 36.8 \text{ N downwards}$$

Some interesting aspects of problem-solving arise from Worked Example 2.4E. If it was assumed that the forces supplied by both the pillars were upwards, this would prove to be wrong. Because care was taken with the directions of the known forces, the correct directions for the forces emerged. Some common sense will always help in these situations. If the beam was pivoted at F_1, and there is a force due to gravity pulling the beam clockwise, then in order to balance the clockwise torque there must be an anticlockwise torque or an upwards force at $\boldsymbol{F}_2$.

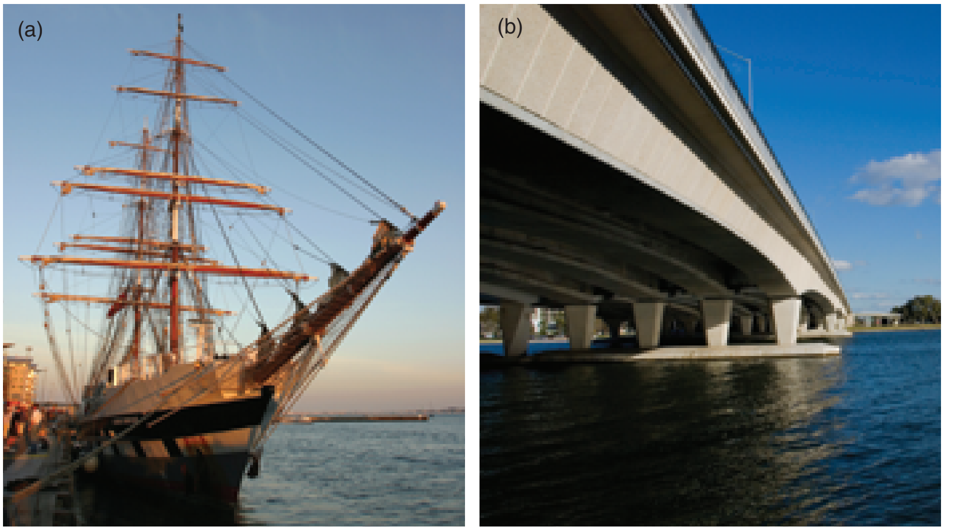


Figure 2.31
(a) The bowsprit projecting from the prow of a sailing ship is a cantilever, and the masts have many vertical cantilevers. (b) The Narrows Bridge is a cantilever bridge spanning the Swan River in Perth.

If there must be a downwards force at $\boldsymbol{F}_1$, then clearly, the beam needs to be attached to pillar 1, using nails, screws, straps or bolts to hold it down. It is worth noting also that if the right-hand support had been moved so that the beam's centre of gravity lay between the two supports, the direction of $\boldsymbol{F}_1$ would change so that both supporting forces would be upwards. You might like to redo the problem using an overhang of 6.0 m. You could also consider what might happen to the value of the supporting forces if a person were to walk from the region between the pillars onto the overhanging span.

Struts and ties

As well as the main beams and pillars, many structures have additional members that help to strengthen them. A structure may be supported by *struts* and *ties*. A strut will be under compression and must be rigid. A tie may be rigid or flexible and will be under tension.

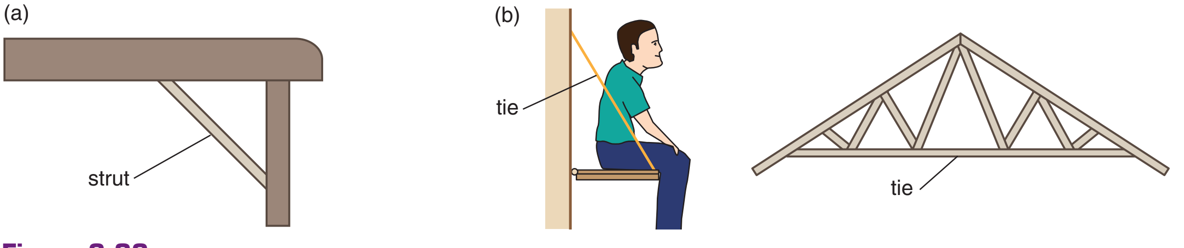


Figure 2.32
(a) A strut helps to support a cantilevered beam and is under compression. (b) A tie helps to support a fold-out bench and is under tension.

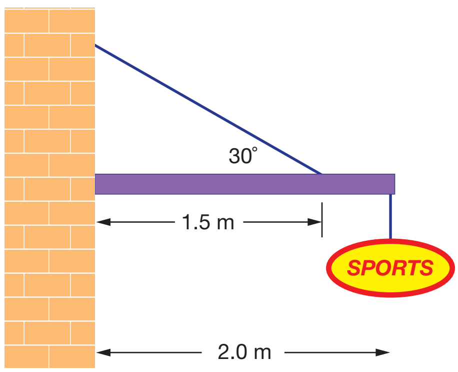
✓ Worked Example 2.4F

A sign of mass 10 kg is suspended from the end of a uniform 2.0 m long cantilevered beam. The beam has a mass of 25 kg and is further supported by a wire tie that makes an angle of 30° to the beam. The wire is attached to the beam at a point 1.5 m from the wall. Use $\boldsymbol{g} = -9.80\ \text{N kg}^{-1}$ and ignore the mass of the wire for these calculations. Find the tension in the wire that is supporting the beam.

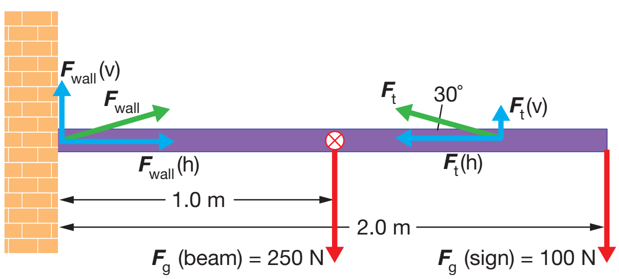
Solution

This is a more complex situation than Worked Example 2.4E. The forces are not simply perpendicular to the beam, so components must be used to determine the torques.

Begin by identifying all the forces acting on the beam in a free-body force diagram. There are four forces to consider. The force due to gravity on the beam, $\boldsymbol{F}_g$(beam), will act at the centre of gravity, 1.0 m from the wall. The tension from the tie wire $\boldsymbol{F}_t$ acts along the wire. The force due to gravity on the sign, $\boldsymbol{F}_g$(sign), pulls downwards at the very end of the beam. The force that the wall exerts on the beam, $\boldsymbol{F}_{wall}$, is not so obvious; in fact we know less about this force than the others as we don't know its magnitude or its direction. If we pay careful attention to our sign convention then the directions of the components of this force will be given in the answer to the calculations. However, let us take an educated guess about the direction of this force. What would happen to the beam if this force was not present? If the wall collapsed or vanished at the point of contact, and the beam was to pivot at the point at which the tie wire is attached to the beam, the beam would probably rotate anticlockwise and swing to the left. To oppose this movement the wall must exert a force that is acting to the right and upwards.



If the point at which the beam meets the wall is considered to be the pivot, then the weight of the beam and the sign supply clockwise torques, and the tension must supply an anticlockwise torque. So first we calculate the perpendicular distance from the line of action of the tension force to the pivot point:



$$r_{\perp t} = 1.5 \times \sin 30°$$

$$r_{\perp t} = 0.75 \text{ m}$$

$m_{beam} = 25$ kg

$r_{\perp \text{ com beam}} = 1.00$ m

$r_{\perp t} = 0.75$ m

$m_{sign} = 10.0$ kg

$r_{sign} = 2.00$ m

$\boldsymbol{g} = 9.80 \text{ N kg}^{-1}$

$$\Sigma\tau_{cw} = \Sigma\tau_{cw}$$

$$F_{beam} r_{\perp \text{ com beam}} + F_{sign} r_{\perp \text{ sign}} = F_t r_{\perp t}$$

$$F_t = \frac{F_{beam} r_{\perp \text{ com beam}} + F_{sign} r_{\perp \text{ sign}}}{r_{\perp t}}$$

$$= \frac{(25.0)(9.80)(1.00) + (10.0)(9.80)(2.00)}{0.75}$$

$$= 588 \text{ N along wire}$$

Physics in action — Structures through the ages

About 4600 years ago the early Egyptians became the first builders of significant stone monuments, in which they housed the bodies of their dead pharaohs. But the ancient Greeks were the first to devise ways of constructing large buildings for public use. Their architecture employed a beam or lintel of stone to span the space between the upright supports. In those times, stone was used for building because of its durability, although timber was used for roofing.

For the ancients, stone added grandeur and a sense of permanence to the structure, but it has two significant drawbacks: it is very heavy and it is weak under tension, even though it is very strong under compression.

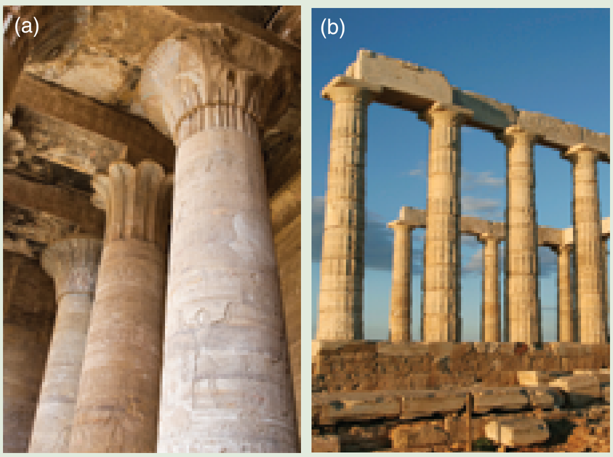


Figure 2.33

(a) The ceiling of the temple at Esna in Egypt has survived remarkably well, but it has been supported by a very large number of columns. Built in the Greco-Roman period (around 500 BCE), the columns have been fashioned to look like bunches of papyrus stalks. (b) The marble temple of Poseidon at Cape Sunium, built in the late fifth century BCE south-east of Athens in Greece, has not survived so well. All that remains of this famous temple are some columns and a few lintels. The columns here are also close together owing to the weakness of stone under tension.

Stone is most suitable as a column to support a lintel beam, since the weight of the lintel is vertically downwards, placing the column under compression. As a lintel, however, stone's weight causes it to sag, and tension cracks can develop at weak points along its lower face. As a consequence, many stone lintels collapsed and the space within the building had to be cluttered with very many supporting columns, causing the floors of such buildings to be taken up with the pedestals of the columns.

The arch

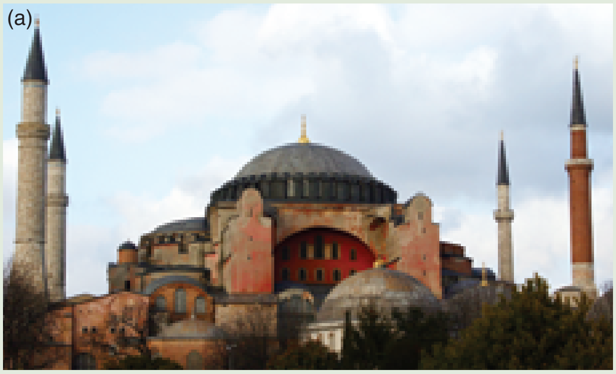
About 500 years after the Greeks, the Romans devised a method for spanning a far greater distance, still using stone. Their solution was to construct a *semicircular* arch using stones shaped by hand and arranged around a wooden form. Once the final stone (the keystone) was put in place, the form could be removed. The great advantage was that every stone in the arch experienced only compression. The results of this innovation were graceful bridges and aqueducts, many of which survive intact today.

Figure 2.34

The aqueduct at Pont du Gard in the south of France. Built by the Romans in about 18 CE, the arches carry water from a spring in the mountains to a settlement some distance away. Building arches within a valley enables the horizontal forces created by each arch to be eventually balanced by forces from the hills on either side.

The weight of an arch does produce one problem, however. When a lintel rests upon a column, its weight is balanced by the upward reaction force provided by the column. There are no horizontal forces to consider. Within a stable arch, however, a horizontal force acting inwards on either side of the arch is required to balance the outward *thrust* caused by the arch. In other words, an inward horizontal force is needed to balance the horizontal component of the thrust of the structure, which acts outwards from the keystone. If this horizontal force is not available, the supports of the arch will be pushed outwards and the arch will collapse. One way that the Romans endeavoured to provide this horizontal force was to build a heavy wall above the arch. The extra weight of the wall acting downward had the effect of enabling the wall to withstand a greater horizontal outward force from the arch.

The Romans also realised that it was possible to build the arch in three dimensions, creating a dome. The Pantheon in Rome, completed in 9 CE, stands as a testament to the skill of its builders. As discussed in the chapter opening, the Hagia Sophia was constructed in the sixth century in Constantinople (Istanbul in Turkey). This church consisted of four piers on which four semicircular arches rested. Resting on the arches was a large



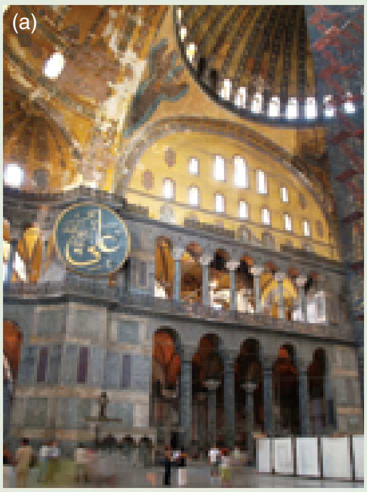
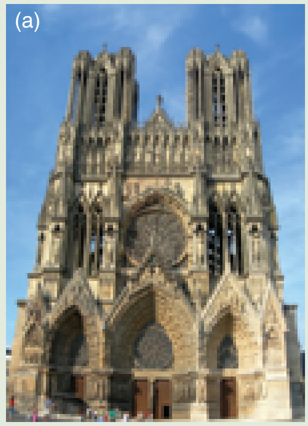


Figure 2.35

(a) On the exterior of the Hagia Sophia are a series of semicircular domes that support the eastern and western walls of the building. A massive rectangular buttress (out of sight) is used on the northern and southern walls.
(b) The interior of the Hagia Sophia. Four piers in a square support four arches and the dome, which is clearly visible, rests on the arches.

Figure 2.36

Battle Abbey, at Hastings in England, was built in the 11th century. These rooms demonstrate that far greater floor space and light within a building can be achieved using arches based on the Roman arch.



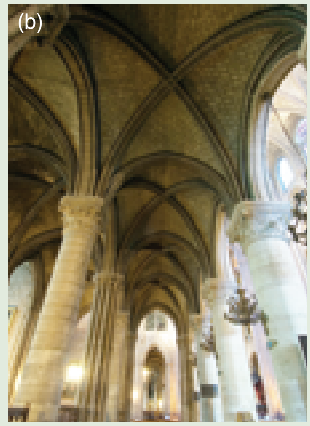


Figure 2.37

(a) The construction of Rheims Cathedral in France, began in 1210. (b) The pointed Gothic arch is a feature of many of the cathedrals built in Europe around this time. The horizontal forces created by the stone in the arch are not balanced by a wall, but are transferred to the outside of the building where flying buttresses accommodate them. (c) The pinnacle on top of each buttress increases the vertical forces through the buttress to ensure that the combined thrust (horizontal thrust from the arch and weight of the buttress and pediment) lies within the structure. This enables the structure to remain stable. In one sense, the cathedral can be considered as a shell suspended within a buttress framework.

dome. One difficulty in the construction of the Hagia Sophia was that the thrust of the dome pushed outwards on the piers. To balance these forces, the walls had to be buttressed with half domes, which in turn were buttressed with quarter domes and so on until the outward thrust force was transmitted to the ground. In 1453 the armies of Sultan Mohammed II took Constantinople and the church was converted to a mosque. In recent times, minarets (towers) were added (Figure 2.35b).

For a Roman arch to remain in place, a horizontal force must be supplied to the base of the arch. Eventually builders found that an arch could be made more stable if it were taller. This so-called Gothic arch was widely used from the end of the 12th century in Europe, when about 80 cathedrals were built. Instead of having to use half domes and heavy walls as the Romans had done, these medieval builders could buttress the Gothic arch with far less weight. This is because a taller arch requires a smaller horizontal reaction force.

As with the Hagia Sophia, the horizontal thrust forces are transferred outside the building, but since the force required is smaller, less massive arches can be used. These supporting arches are called flying buttresses. This form of buttressing was popular for its aesthetic appeal and because it allowed a great deal of light into the building. By adding extra weight to the top of a buttress—usually in the form of a statue or spire—the horizontal forces from the arches could be supplied even more easily.

Arch bridges

The Romans used the arch to great effect both as a bridge and as a structural unit when building aqueducts. But apart from stone and wood, no new materials were available to bridge-builders until the late 18th century, when iron began to be smelted in sufficient quantities to make it economical. Initially cast iron was used. Like stone, cast iron is strong under compression and weak under tension. The first iron bridge—built over the River Severn at Ironbridge, England, in 1779—used short iron struts to follow the design of a stone arch bridge. As with a stone arch, a metal arch bridge carries its load by placing its members under compression. For an arch bridge, the arch can be located above or below the road line.

A modern arch bridge closer to home is the Gladesville Bridge in Sydney. Using reinforced and prestressed concrete, this bridge mimics a Roman arch. The deck of the bridge spans 305 m across the Parramatta River. The Sydney Harbour Bridge, opened in 1932 with a span of 503 m, is also an arch bridge, but the arch is made from steel trusses from which the deck is suspended.

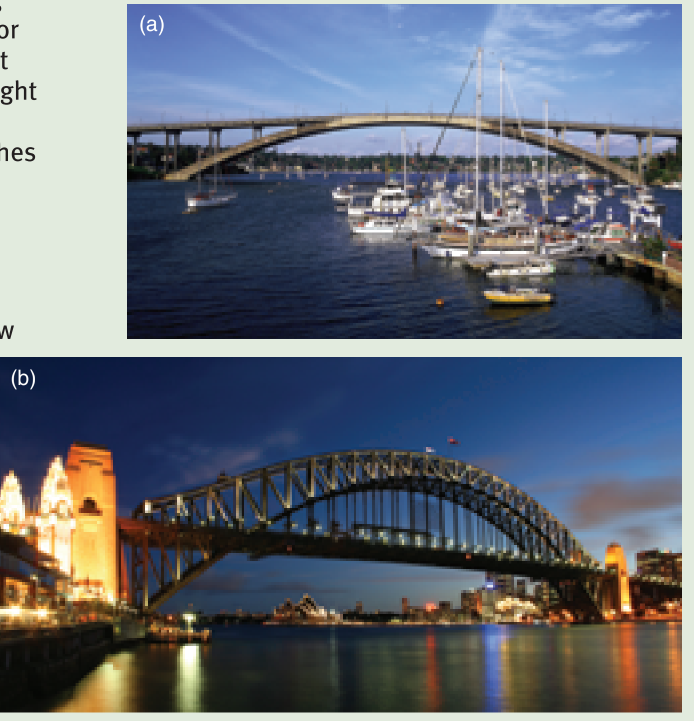


Figure 2.38

(a) The Gladesville Bridge across the Parramatta River. (b) The Sydney Harbour Bridge. Each bridge uses an arch, but the designs produce different effects.

2.4 SUMMARY Equilibrium

- A body is in translational equilibrium when the sum of the forces acting on it is zero, i.e. $\Sigma \boldsymbol{F} = 0$.
- If the sum of the forces acting on a body is zero, the sum of the individual components of the forces are also zero, i.e. $\Sigma \boldsymbol{F}_x = 0$, $\Sigma \boldsymbol{F}_y = 0$ and $\Sigma \boldsymbol{F}_z = 0$.
- Where forces act in different directions at different points on a body, the forces act as a couple, and an unbalanced torque may occur even though the components of the forces are equal in magnitude.
- For a body or system in rotational equilibrium, the sum of all the torques acting must be zero, i.e. $\Sigma\tau = 0$.
- For a body or system to be in static equilibrium, it must be in translational and rotational equilibrium, i.e. $\Sigma \boldsymbol{F} = 0$ and $\Sigma\tau = 0$. As a consequence of the translational equilibrium condition, $\Sigma \boldsymbol{F} = 0$, $\Sigma \boldsymbol{F}_x = 0$ and $\Sigma \boldsymbol{F}_y = 0$ must also be true.

2.4 Questions

Use $g = -9.80$ N kg^{-1} when answering these questions.

1 Which of the following are in translational equilibrium?
A A stationary elevator
B An elevator going up with constant velocity
C An aeroplane during take-off
D A container ship sailing with constant velocity
E A car plummeting off a cliff

2 Two window-cleaners work on a platform that is supported by four cables. The platform has a mass of 50 kg, and the cleaners weigh 600 N and 850 N. Assuming that all the weight is evenly distributed, calculate the tension in each of the cables.

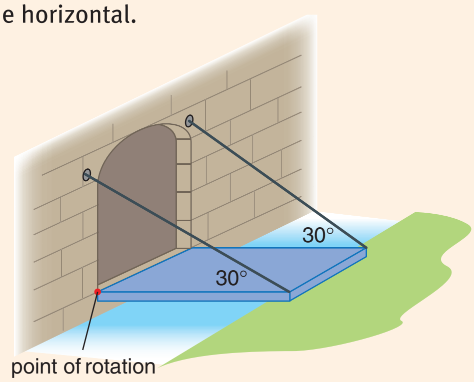
3 A bridge over a river is made from steel girders that have a total mass of 5000 kg. The bridge is supported by two pillars which each support half of the load. The bridge is designed to support a further load of 20 tonnes and has a safety factor of 8 (i.e. it can support eight times the designed maximum load.) What force must each of the two supporting pillars be capable of providing?

4 A 60 kg tight-rope walker carries a long beam with a mass of 30 kg across a 10 m long wire. When she is at the centre of the wire (i.e. 5 m across), each section of the wire makes an angle of 5° to the horizontal. Assuming that the mass of the wire is negligible, calculate the tension within it.

5 Two children are balanced on a uniform see-saw which is supported in the middle on a pivot. One child weighs 200 N and is 1.2 m from the pivot, while the other child is seated 1.5 m from the pivot. What is the weight of the second child?

6 A uniform 2.0 kg horizontal beam 50 cm long is bolted to a brick wall and supports a 5.0 kg lighting fixture. Calculate the torque produced by the combined weight of the beam and the light about the point where the beam meets the wall.

7 A 10 m long drawbridge is supported by two cables which extend from two holes either side of a door in a castle wall. The bridge has a mass of 700 kg and the tension is the same in both cables. The bridge is just about to touch the ground and the cables make an angle of 30° to the horizontal.

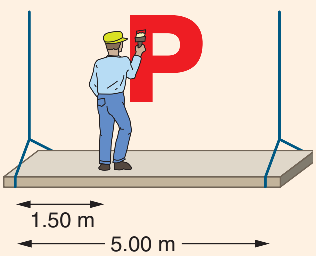


a Write an expression for the torque produced by each of the tensile forces that act around the axis of rotation of the bridge.
b Write an expression for the horizontal and vertical components of the tension force.
c Calculate the size of the tension in each cable.

8 A train engine passes over a 20 m bridge span which is supported by two columns X and Y. The engine has a total mass of 5.0 tonnes. At one instant column Y produces a reaction force of 30.6 kN. If the spanning beam is uniform and has a mass of 5.0 tonnes, where is the centre of mass of the train?

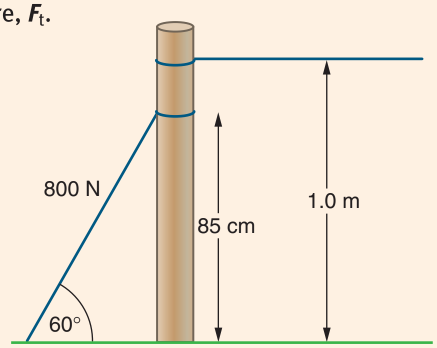
9 A ladder of length 4.8 m and mass 16 kg is leaning against a wall so that it makes an angle of 65° to the horizontal. Calculate the magnitude of the torque exerted on the ladder (taken around where it contacts the ground) by each of the following forces:
a the weight of the ladder
b the weight of a person of mass 50 kg standing one-quarter of the way up the ladder
c the weight of a person of mass 50 kg standing three-quarters of the way up the ladder.

10 a A 5.00 m long painter's platform has a mass of 20 kg and is supported by two ropes as shown. A 70 kg painter stands 1.50 m from the left. Calculate the tension in each supporting rope.



b The painter alters his position so that the left-hand rope now experiences a tension of 557 N and the other a tension of 325 N. Where is the painter now standing in relation to the left-hand rope?

11 The end-post of a wire fence is held in position by a backstay which is under a tension of 800 N at an angle of 60° to the horizontal. The geometry of the situation is shown in the diagram.
a Determine values for the horizontal and vertical components of the tension in the backstay wire.
b By considering the base of the post to be a pivot point, determine the size of the tension in the fence wire, F_t.



Assume that the universal constant of gravitation, $G = 6.67 \times 10^{-11}$ N m^2 kg^{-2}; gravitational field strength on surface of the Earth, $g = 9.80$ N kg^{-1}.

1 The gravitational force of attraction between Saturn and Dione, a satellite of Saturn, is equal to 2.79×10^{20} N. Calculate the orbital radius of Dione. Data: mass of Dione = 1.05×10^{21} kg, mass of Saturn = 5.69×10^{26} kg.

2 A person standing on the surface of the Earth experiences a gravitational force of 935 N. What gravitational force will this person experience at a height of 2 Earth radii above the Earth's surface?
- **A** 935 N
- **B** 234 N
- **C** zero
- **D** 104 N

3 During a space mission, an astronaut of mass 80.0 kg initially accelerates at 30 m s^{-2} upwards, then travels in a stable circular orbit at an altitude where the gravitational field strength is 8.20 N kg^{-1}.

a What is the apparent weight of the astronaut during lift-off?
- **A** zero
- **B** 780 N
- **C** 2400 N
- **D** 3200 N

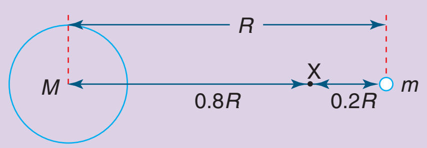
b During the lift-off phase, the astronaut will feel:
- **A** lighter than usual
- **B** heavier than usual
- **C** the same as usual.

c The weight of the astronaut during the lift-off phase is:
- **A** lower than usual
- **B** greater than usual
- **C** the same as usual.

d During the orbit phase, the apparent weight of the astronaut is:
- **A** zero
- **B** 780 N
- **C** 2400 N
- **D** 660 N

e During the orbit phase, the weight of the astronaut is:
- **A** zero
- **B** 780 N
- **C** 2400 N
- **D** 660 N

4 Two stars of masses *M* and *m* are in orbit around each other. As shown in the following diagram, they are a distance *R* apart. A spacecraft located at point X experiences zero net gravitational force from these stars. Calculate the value of the ratio *M*/*m*.



5 Neptune has a planetary radius of 2.48×10^7 m and a mass of 1.02×10^{26} kg.

a Calculate the gravitational field strength on the surface of Neptune.

b A 250 kg lump of ice is falling directly towards Neptune. What is its acceleration as it nears the surface of Neptune? Ignore any drag effects.
- **A** 9.8 m s^{-2}
- **B** zero
- **C** 11 m s^{-2}
- **D** 1.6 m s^{-2}

6 Given that the mass of the Earth is 5.98×10^{24} kg and the mean distance from the Earth to the Moon is 3.84×10^8 m, calculate the orbital period of the Moon. Express your answer in days.

7 One of Jupiter's moons, Leda, has an orbital radius of 1.10×10^{10} m. The mass of Jupiter is equal to 1.90×10^{27} kg. Calculate the:
- **a** orbital speed of Leda
- **b** orbital acceleration of Leda
- **c** orbital period of Leda (in days).

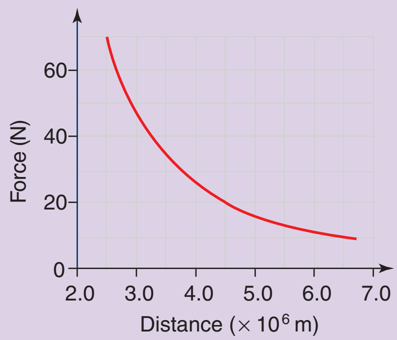
8 **a** Which of the following best explains what is meant by a satellite being in a geosynchronous orbit?
- **A** It is orbiting the Earth.
- **B** It is orbiting the Moon and remains above the same location.
- **C** It is orbiting the Earth and remains above the same location.
- **D** It is orbiting the Earth and returns to the same location every 24 hours.

b What is the purpose of such an orbit?

c Assuming that the length of a day on Earth is exactly 24 hours, calculate the radius of orbit of a geostationary satellite (mass of Earth = 5.98×10^{24} kg).

9 The planet Mercury has a mass of 3.30×10^{23} kg. Its period of rotation about its axis is equal to 5.07×10^6 s. For a satellite to be in a synchronous orbit around Mercury, calculate:
- **a** the orbital radius of the satellite
- **b** its orbital speed
- **c** its orbital acceleration.

10 The following graph shows the force on a 20 kg rock as a function of its distance from the centre of the planet Mercury. The radius of Mercury is 2.4×10^6 m.



A 20.0 kg rock is speeding towards Mercury. When the rock is 3.00×10^6 m from the centre of the planet, its speed is estimated at 1.00 km s^{-1}. Using the graph, estimate the:
- **a** increase in kinetic energy of the rock as it moves to a point that is just 2.50×10^6 m from the centre of Mercury
- **b** kinetic energy of the rock at this closer point
- **c** speed of the rock at this point
- **d** gravitational field strength at 2.50×10^6 m from the centre of Mercury.

11 Two satellites S_1 and S_2 are in circular orbits around the Earth. Their respective orbital radii are R and 2R. The mass of S_1 is twice that of S_2. Calculate the value of the following ratios and use the following answer key: **A** 1, **B** $\sqrt{2}$, **C** $1/\sqrt{8}$, **D** 4.
- **a** orbital period of S_1/orbital period of S_2
- **b** orbital speed of S_1/orbital speed of S_2
- **c** acceleration of S_1/acceleration of S_2.

12 Earth is in orbit around the Sun. The Earth has an orbital radius of 1.50×10^{11} m and an orbital period of 1 year. Use this information to calculate the mass of the Sun.

The following information relates to questions 13–15.

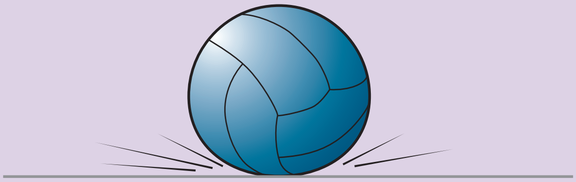
Charon is the largest of Pluto's moons. It orbits Pluto with a period of 6.40 days and an orbital radius of 19 600 km. Nix, another of Pluto's moons, was discovered using the Hubble Space Telescope in 2005. Nix has an orbital radius of 49 000 km—about double that of Charon.

13 Which of the following statements is correct?
- **A** The gravitational force between Nix and Pluto is greater than the force between Charon and Pluto.
- **B** The gravitational force between Nix and Pluto is less than the force between Charon and Pluto.
- **C** The gravitational force between Nix and Pluto is equal to the force between Charon and Pluto.
- **D** The gravitational forces cannot be compared with the information given.

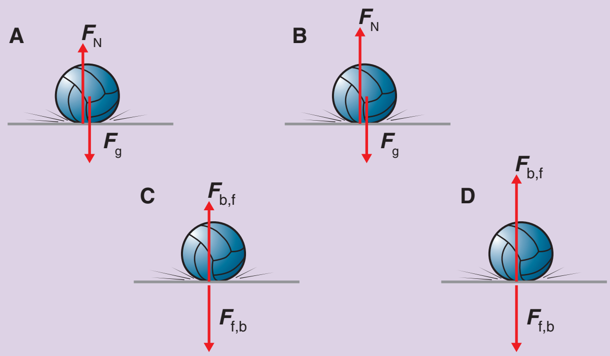
14 Use Kepler's third law to determine the orbital period of Nix (in days).
15 Use the data relating to Charon to calculate the mass of Pluto.

The following information applies to questions 16–18.
The International Space Station orbits Earth at an altitude of 380 km. The Optus D2 satellite has a much higher orbit at an altitude of 36 000 km. Assume that the radius of Earth is 6.37×10^6 m, and the mass of Earth is 5.98×10^{24} kg.

16 Determine the value of the ratio: speed of ISS/speed of Optus D2.
17 Determine the value of the ratio: period of ISS/period of Optus D2.
18 Determine the value of the ratio: acceleration of ISS/acceleration of Optus D2.
19 On the diagram below, draw and identify the forces that are acting on the ball at this instant, being careful to show the relative sizes of the forces.



20 **a** Which of the following correctly represents the action/reaction forces acting between the ball and the floor at this instant? (One or more answers.)



b Explain your answer to part a.

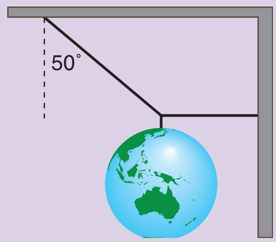
21 The contents of a wheelbarrow produce a clockwise torque of 445 N m about the wheel axle, which is 1.60 m from the end of its handles. If this wheelbarrow is to be wheeled away, what total force must be exerted on the handles? Assume the force acts at 90.0° to the handles.
22 Building wreckers wish to knock over a concrete wall. They plan to use a ball and chain which exerts 5.00×10^3 N to hit the wall at a point that is 3.00 m above the ground. What torque is developed on the wall around the base? In what ways could the wrecker increase the torque exerted on the wall?

The following information applies to questions 23 and 24.
A 6.00 m ladder of mass 7.50 kg leans against a wall at a 65.0° angle to the horizontal.

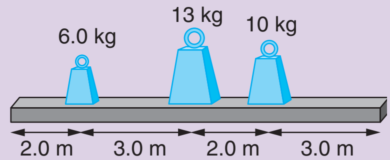
23 What is the torque (taken around where the ladder rests on the ground) exerted on the ladder by the weight of the ladder itself?
A 220 N m
B 22 N m
C 31 N m
D 186 N m
E 93.2 N m
24 A 60.0 kg woman is standing one-third of the way up the ladder. Calculate the torque (taken around where the ladder rests on the ground) exerted on the ladder by the woman?

The following information applies to questions 25 and 26.
A crane is lifting a prefabricated concrete wall of mass 4.50 tonnes. Assume that a single steel cable is being used to lift the load, and ignore the mass of the cable in your calculations.

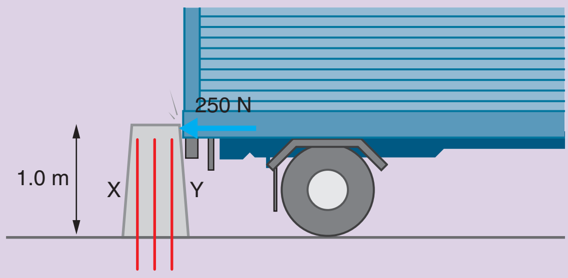
25 What is the magnitude of the tensile force acting in the cable when the load is being held stationary above the ground?
A 44 N
B 44 kN
C 4.5 N
D zero
E 9.8 N
26 The load is now lifted at a constant speed of 2.00 m s^{-1}. Which of the following best describes how the tension acting in the cable compares with the tension value in Question 25?
A The tension is now greater.
B The tension in the cable is equal to that in Question 25.
C The tension in the cable is less than that in Question 25.
D This cannot be determined from the information given.
27 A 1.50 kg model of the Earth is suspended by two long wires in a school library as shown. Calculate the tension in each wire.



28 A 2.00 kg beam supports masses of 6.00, 10.0 and 13.0 kg at the positions shown. The beam is resting on supports at each end. Calculate the magnitudes of the support forces that are acting at each end of the beam.



29 A crane with a horizontal arm is lifting a steel girder of mass 1.50 tonnes. Initially, the load was being lifted at the far end of the horizontal arm, a distance of 25.0 m from the pivot of the crane. Ignore the mass of the cable when answering these questions.
a The crane carries a counterweight of mass 20.0 tonnes. How far from the pivot should this counterweight be positioned so that its torque balances the torque of the load?
b The driver decided that it would be wise to bring the load in closer to the pivot. What is the benefit of bringing the load in closer to the crane pivot?
30 A barrier 1.00 m high in an underground carpark is a composite material made of concrete and steel reinforcing rods. A truck reverses into the barrier and exerts a force of 250.0 N on it, as shown in the diagram.



a Calculate the magnitude of the torque around the base of the barrier that this force creates.
b Is the concrete more likely to crack at X or Y as a result of this collision? Explain.

3 Understanding electromagnetism

Less than 200 years ago most people thought of electricity and magnetism as being quite separate, distinct phenomena, with little but curiosity value. But in 1820 Hans Christian Oersted discovered that an electric current could produce a magnetic field. Once that link was discovered, the knowledge of electromagnetism and our uses for it increased at a tremendous rate, laying the foundations for our modern way of life. The progress that has occurred as a result of our understanding of the connection between magnetism and electricity, gained only in the last 200 years, has been one of the truly remarkable achievements of humankind.

Our modern craving for electric power, however, has also created one of the 21st century's greatest challenges: to find sustainable ways of generating the huge amounts of electrical energy needed to power modern technological societies. The understanding of the basic physics of electric power you will gain from this chapter is an essential first step in meeting that challenge.

This photo (opposite) shows an aurora, one of nature's most beautiful phenomena. Auroras are produced when electrically charged particles from the Sun enter the Earth's magnetic field.

Extra material that revisits and expands on 'Understanding electromagnetism' can be found on the *ePhysics* CD.

By the end of this chapter

you will have covered material from the study of understanding electromagnetism, including:

- the nature of current, voltage and electric circuits
- the nature of magnets and the origin of magnetic fields
- magnetic forces on currents and moving charges
- applications of magnetic forces, including electric motors.

3.1 Magnetic fields

Simple magnetism

If you put a paper clip or pin near a magnet, it will experience a force and be pulled towards the magnet. The space around the magnet must therefore be affected by the presence of the magnet; that is, there is a **magnetic field** around the magnet. An even clearer way to see the presence of this field is to sprinkle iron filings on a piece of card held over a magnet. The iron filings line themselves up with the field, making it clear that there are 'lines of force' which seem to run from one end of the magnet to the other.

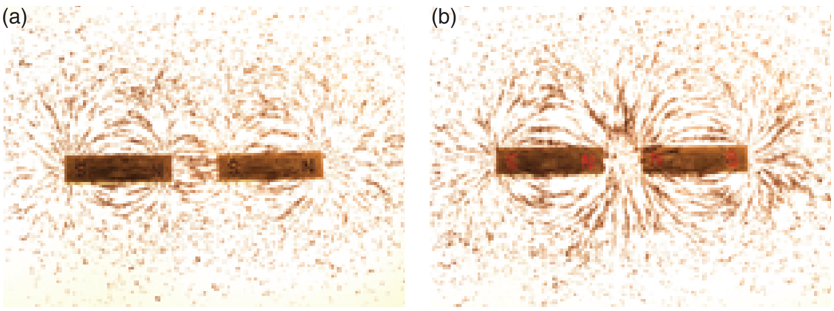


Figure 3.1

Iron filings sprinkled around magnets (a) with unlike poles together and (b) with like poles together. The patterns in the fields clearly show the attraction and repulsion.

The most obvious magnetic effect is that any piece of iron will be attracted to a magnet. On experimenting with two magnets, however, we find that each end of a magnet acts differently from the other. For example, if one end of a magnet is attracted by a second magnet, then the other end will be repelled. We refer to the ends of magnets as **magnetic poles**.

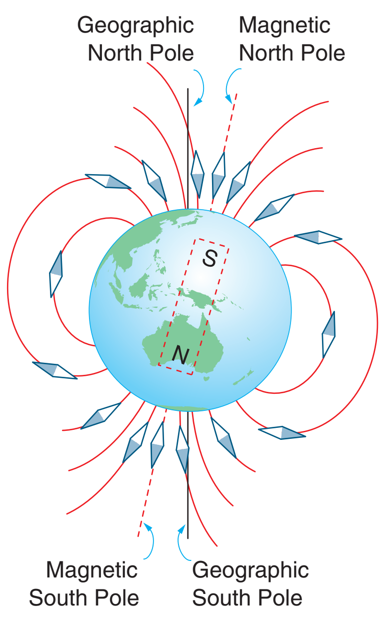


Figure 3.2

The Earth acts as though it has a south magnetic pole near the geographic north pole! The 'magnetic north pole' is the place to which the north end of a compass appears to point.

> Like magnetic poles ***REPEL*** each other; unlike magnetic poles ***ATTRACT*** each other.

A magnet suspended so that it is free to rotate horizontally will always align itself in a north–south direction. This is the reason the poles were labelled *north pole* and *south pole*. However, one must be careful to distinguish between the words 'north' and 'south' when used in a magnetic and a geographic context (see Figure 3.2).

If the magnet is free to swing vertically as well, in the southern hemisphere the north pole end will point upwards as well as northwards. In the northern hemisphere, the north end points downwards. It is as though the Earth itself is acting as a huge magnet, with its south pole to the geographic north and its north pole to the geographic south.

The properties of magnets may remind us of the forces between electrical charges, where like charges repel and opposite charges attract and the force of attraction or repulsion increases as the distance between the charges decreases. This is one reason why some 19th century philosophers believed that there might be some connection between magnetism and electricity. On the other hand, there are some clear differences too. Magnets are more or less 'permanent', whereas it is hard to keep an electric charge on an object, such as a rubbed plastic comb, for much more than 10 or 15 minutes. Magnetic poles do not run away through metal wires to ground as an electric charge will do.

Can we obtain separate north and south poles?

We could try cutting a magnet in half, but all we get is two smaller magnets, each with its own north and south poles. No matter how often we keep cutting magnets, we always get more little magnets with two opposite poles.

Originally it was thought that the poles were some sort of physical entity which gave rise to the lines of force, and that they were embedded near the ends of the magnet. Eventually it became clear, particularly from the fact that cutting a magnet in half seemed to produce two new poles, that there was no such thing as a separate pole, but that the lines of force continued right through the magnet and out the other end. Because magnets always have two poles, they are said to be *dipolar*.

Magnets are ***DIPOLAR*** and the field around a magnet is called a ***DIPOLE FIELD***.

Physics file

The north end of the needle in a good compass made for use in the southern hemisphere is a little heavier than the south end to compensate for the upward pointing field. In the northern hemisphere, the opposite is the case. So if a compass made for use in one hemisphere is taken to the other, it will appear to be a little out of balance.

Why does a magnet attract an unmagnetised piece of iron?

In experimenting with magnets and iron, one soon discovers that while all types of iron become magnetised when placed near a magnet, some types of iron lose their magnetism once the magnet is removed and others don't. The first type is called 'soft' iron and the second 'hard' iron. (Nails are made of soft iron; chisels are hard iron—the difference is in the tempering process.) We can place a piece of soft iron close to a permanent magnet and use iron filings to look at the fields around the magnet and soft iron. The filings show us that in the presence of the permanent magnet, the piece of iron has also become a magnet. The permanent magnet has somehow 'induced' the piece of iron to become a magnet. We call this effect *induced magnetism*. We can see, therefore, why a permanent magnet will attract another piece of iron. It induces the other piece of iron to become a magnet with the opposite pole closest, and thus they attract. Induced magnetism is a temporary phenomenon. As soon as the permanent magnet is removed, the soft iron loses its magnetism.

Figure 3.3
Magnets are always dipolar. When a magnet is broken, two new poles appear at the broken ends.

MAGNETIC FIELD LINES are represented as arrows that flow out from the north pole of a magnet into the surrounding space and return to the south pole. Internally the magnetic field flows from the south pole to the north pole, completing the loop. If a compass was placed at any point in the space around a magnet, its needle would align with the field line.

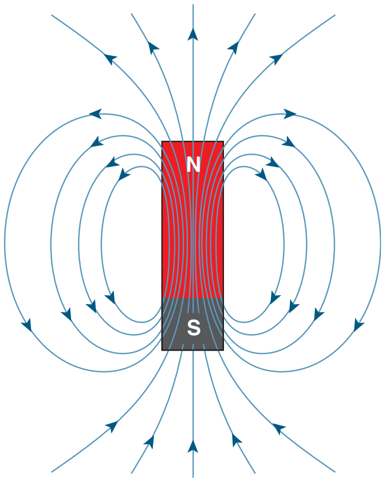


Figure 3.4
The field lines around and inside a bar magnet. The lines show the direction of the force on an (imaginary) single north pole.

Comparing magnetic fields

Before trying to answer fundamental questions about the nature of magnetism, we need to have a clear understanding of the concept of a magnetic field and its use in describing magnetic interactions. First, let us look at how we determine the direction of a magnetic field. We define the direction of the magnetic field as the direction in which the magnetic force is exerted on the north pole of another magnet placed in the field. The force on a south pole, then, is always in the direction opposite to the field direction.

Figure 3.5
A strong magnet in the base induces each little leaf to become a magnet.

Figure 3.7
A compass magnet is free to align itself with the Earth's magnetic field. In a high-quality compass like this one, the case is filled with light oil to dampen the motion of the magnet, and the north end is slightly weighted to balance the upward component of the Earth's field in the southern hemisphere.

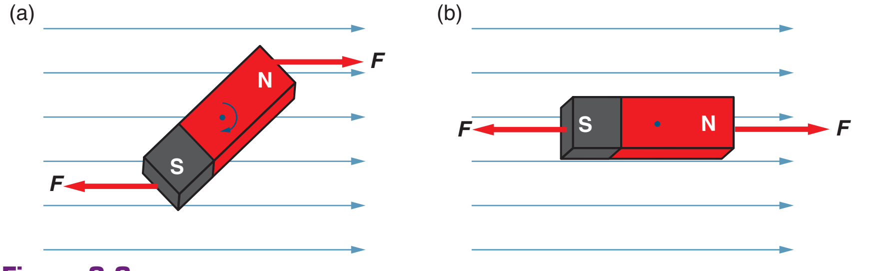


Figure 3.6
The torque (turning force) on a magnet in a magnetic field. (a) The torque will tend to rotate the magnet. (b) The magnet is still subject to the opposing forces but now experiences zero torque.

Physics in action — Natural magnets

Magnetite, a natural oxide of iron (Fe_3O_4), is moderately abundant on the Earth's surface. Occasionally, natural pieces of magnetite are found to be permanently magnetised, perhaps as a result of lightning strikes. These natural magnets are known as *lodestones*. Early humans were probably aware of the properties of these natural curiosities, but the first recorded use of a lodestone as a compass was by a Chinese emperor around 2600 BCE. He is said to have had a compass, in the form of the figure of a woman who always pointed south, mounted on the front of his chariot. It appears that the Chinese also discovered, around 1100, that a steel needle could be magnetised, and then used as a compass, by stroking it with a lodestone.

Around 600 BCE the Greeks also discovered the properties of lodestones, and even wondered about their connection with the electrical phenomena associated with amber rubbed with fur. The Greeks gave us the words *electricity*, meaning 'amber', and *magnet*, from Magnesia, an area where lodestones were found. (Another version is that the word 'magnet' came from the name of a shepherd, Magnes, who found that the iron on the end of his staff was attracted to certain stones on Mount Ida.)

Describing magnetic flux density with vectors

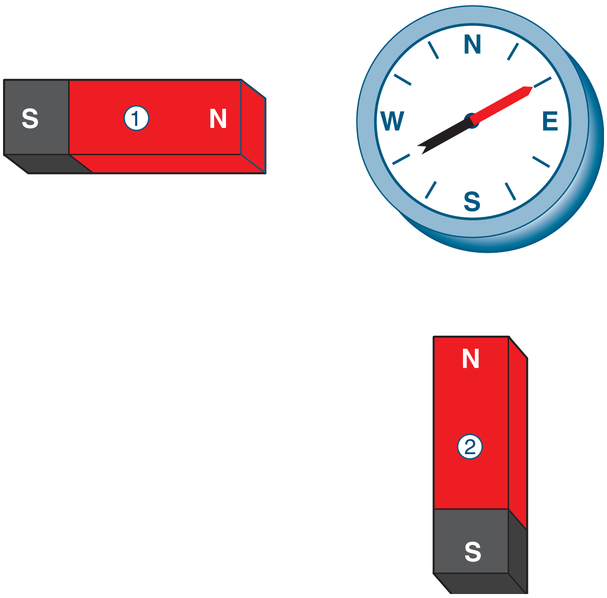


Figure 3.8
The compass needle points away from the north ends of the two magnets, but clearly magnet 1 is stronger than magnet 2 in this case.

Physics file

The magnetic flux density (***B***) is the term used to describe the strength of the magnetic field. It can be likened to the gravitational field strength around a planet. The closer you are to the poles of a magnet, the greater the magnetic flux density. The magnetic field lines around the pole of a magnet are denser than the same field lines further from the pole (see Figure 3.4). This is a convenient way of thinking about magnetic flux density.

Although we are not ready to define the flux density of a magnetic field in absolute terms, we can certainly compare the relative flux densities of two fields. Consider this simple experiment. Place a small compass on a table. Align two magnets so that they are at equal distances from the compass, but at right angles to each other, and with their north ends pointing towards the compass. The north end of the compass will try to point away from both magnets. In Figure 3.8, magnet 1 is clearly stronger than magnet 2, since the north pole of the compass is more strongly repelled by it than by magnet 2. If the magnets were both of equal strength, the compass would make an angle of 45° to the axes of the magnets.

A simple piece of vector analysis enables us to compare the magnetic flux densities of the two fields at any point in the field. The symbol $\boldsymbol{B}$ is used to represent the flux density. In Figure 3.9, if $\boldsymbol{B}_1$ represents the field due to magnet 1 at the location of the compass and $\boldsymbol{B}_2$ represents that of magnet 2 at the same place, we can see that they can be added vectorially to give $\boldsymbol{B}_r$, the resultant field to which the compass will respond.

Strong magnets can quite easily produce magnetic fields of around 10 000 times the strength of the Earth's field. The most intense artificial fields produced are over a million times that of the Earth. But of course these fields are concentrated into a few cubic centimetres or less, while the Earth's field extends not only over the whole Earth, but for many thousands of kilometres out into space!

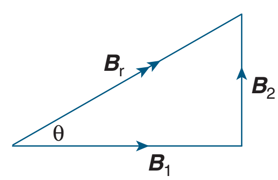


Figure 3.9
The two magnetic fields add as vectors, giving a resultant field to which the compass needle responds.

Physics in action — The strength of the Earth's magnetic field

To get some idea of the strength of the Earth's field compared with that of a typical magnet, we can use the Earth itself in place of one of the two magnets in an experiment similar to the one described in Figure 3.8. The bar magnet is aligned east–west and a number of compasses are positioned along a line extending lengthwise from the magnet. At each position, the direction that the needle points is an indication of the relative strength of the Earth's field compared with that of the magnet.

Where the compass points north-east (i.e. $\theta = 45°$), we know that the two fields are equal in strength. The relative strengths at the other positions can be found from the value of the tangent of the angle ($\tan\theta$), as in the previous example. The graph in Figure 3.10b gives the magnetic field strength of a typical bar magnet at various distances from its end, in terms of the strength of the Earth's field at that location.

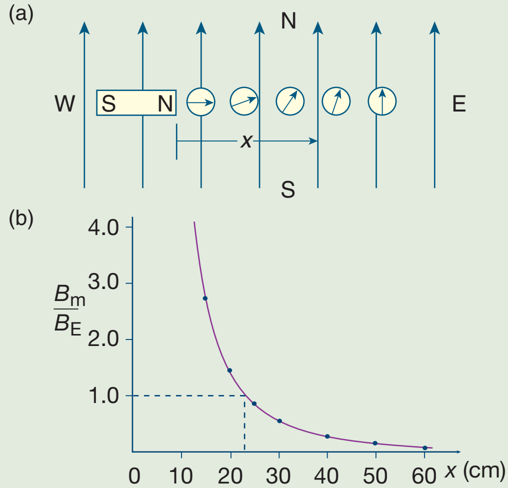


Figure 3.10
(a) The experimental set-up and (b) the graph of B_m/B_E, the ratio of the strength of the field of the magnet compared to the Earth's field, against distance in cm from the magnet. Note that in this case the field of the magnet is equal to the Earth's field at about 23 cm from the magnet.

Figure 3.11
Hans Christian Oersted discovered the magnetic effect of an electric current in 1820.

The foundations of electromagnetism

The first discoveries

Early experimenters wondered whether there were ways (other than using lodestones) to produce a magnetic field. In 1820 the Danish physicist Hans Christian Oersted came up with an answer that was to have enormous consequences for the development of human society.

As a young physics graduate in Copenhagen in the early 1800s, Hans Christian Oersted was looking for a university teaching job, but without much success. To earn a little income he gave public lectures on recent developments in physics—particularly the phenomena associated with an electric current from a 'voltaic pile'. These lectures became so popular that the University of Copenhagen created a special position for him.

Oersted was also something of a philosopher and had visited the famous German philosophers of the time and studied their work. As a result, he had an inclination to believe in the 'unity of nature', the idea that everything is somehow connected. For this reason, he felt that there must be some link between electricity and magnetism, even perhaps that they were different aspects of the same phenomenon. While preparing for a lecture, he noticed that when he switched on a current from a 'voltaic pile', a magnetic compass nearby moved. Further investigation convinced him that it was indeed the current that was affecting the compass.

His discovery can be understood by placing some small compasses near a vertical wire through which a strong electric current is flowing. The compasses tend to become aligned at a tangent to circles around the wire. The current seems to be creating a circular magnetic field. The stronger the current, and the closer the compasses are to it, the greater the effect. The *electric current* is creating a *magnetic field*.

Oersted recognised the significance of this discovery. He went on to find that not only did an electric current create a magnetic field, but that a wire carrying an electric current also experiences a force if it is placed in a magnetic field. Within weeks of hearing of Oersted's work, the Frenchman André-Marie Ampère had given a comprehensive mathematical description of the effects. He also performed new experiments that showed that an electric current responds to the magnetic field produced by another electric current—magnetism without magnets!

CONVENTIONAL CURRENT is a flow of positive charge around a circuit or through space, and ***ELECTRON CURRENT*** is a drift of negative charge in a particular direction. In electromagnetism it is the convention to use the direction of the conventional current as the direction of current flow.

It is easy to remember the direction of the magnetic field around a current by using the **right-hand grip** rule. If you were to grasp the conductor with your right hand in such a way that your thumb points in the direction of the conventional electric current, your fingers curl around in the direction of the field.

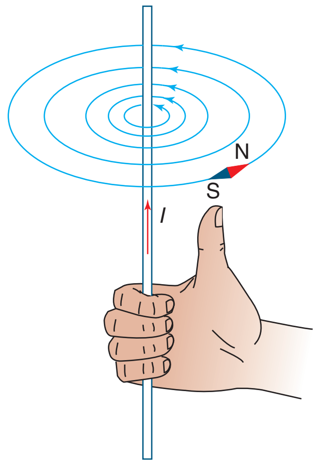


Figure 3.12
The direction of the field around a straight wire is given by the right-hand grip rule.

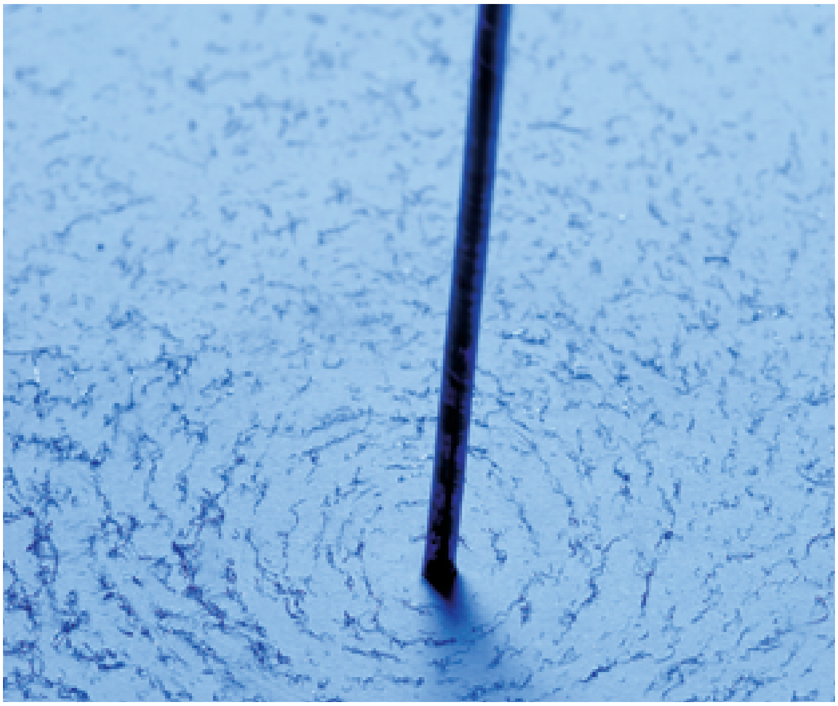


Figure 3.13
The magnetic effect around a current-carrying wire. The iron filings are acting as little compasses and show the circular nature of the magnetic field.

Physics file

Naturally, the closer we are to the electric current, the stronger the magnetic field. It was found that the magnetic field decreased directly with distance: twice as far away, the magnetic field is half as strong. As well, the magnetic field is directly proportional to the electric current: twice the electric current gives twice the magnetic field. Mathematically this can be written as:

$$\boldsymbol{B} = \frac{KI}{r}$$

where $\boldsymbol{B}$ is the magnetic flux density, I is the current, r is the distance from the wire and K is a constant which depends on the medium around the current.

Magnetic fields around currents, magnets and atoms

The great significance of the work of Oersted and Ampère was that magnetism was seen to be primarily an *electrical* phenomenon. Magnetism could be obtained without magnets! Ampère, in fact, showed that the magnetic effects of a coil of wire carrying a current (a *solenoid*) were just the same as that of a permanent magnet of similar size and shape. It is easy to see why this is so.

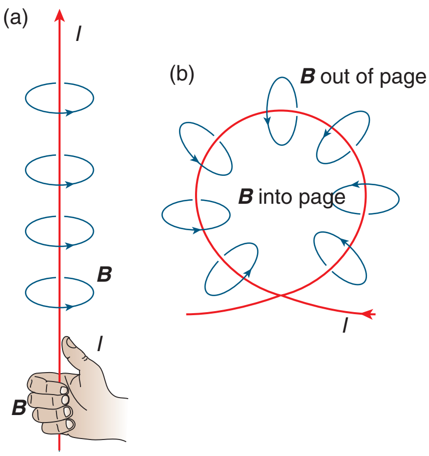


Figure 3.14
(a) The field around a straight current-carrying wire is circular. (b) The wire is bent into a clockwise loop. The field is now into the page in the centre and out of the page around the outside.

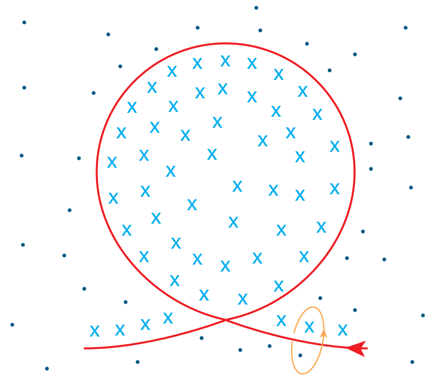


Figure 3.15

Field lines coming out of the page are shown by dots, those going into the page by crosses. The density of the crosses inside the loop will be greater than the density of the dots outside, suggesting a stronger field inside the loop.

If a long wire is made into a circle and a current is passed through it, the magnetic field is enhanced inside the circle and somewhat diminished outside. This is because inside the circle the field from all parts of the wire is pointing in the same direction (into the page if the current is clockwise), whereas at any point outside the circle the contribution to the field from opposite sides of the loop is in opposite directions, and so will tend to cancel.

We often need to draw field lines pointing into or out of the page to represent the magnetic field around a current. These can be represented by dots if the lines are coming out of the page and by little crosses if they are going into the page. (The crosses represent the tail feathers of an arrow going away from you, and the dots represent the point of the arrow coming towards you.) So the field around the loop could be drawn as in Figure 3.15, each cross representing a ring of field going into the page and each dot the ring re-emerging from the page.

If many loops are placed side by side, all their fields add and there is a much stronger effect. This can easily be done by winding many turns of wire into a coil (*solenoid*). The field around the solenoid is like the field around a normal bar magnet, just as Ampère said. This can be checked either by exploring the field with a small compass, or by putting iron filings on a card around the solenoid.

This type of field, in which lines appear to converge towards two 'poles' at either end of the solenoid or bar magnet, is called a *dipole field*. The Earth's field is another example of a dipole field. By way of contrast, the field around a long straight current is a *non-dipole field*—the lines do not converge towards any 'poles' at all.

Practical activity

28 Strength of the magnetic field inside a cell

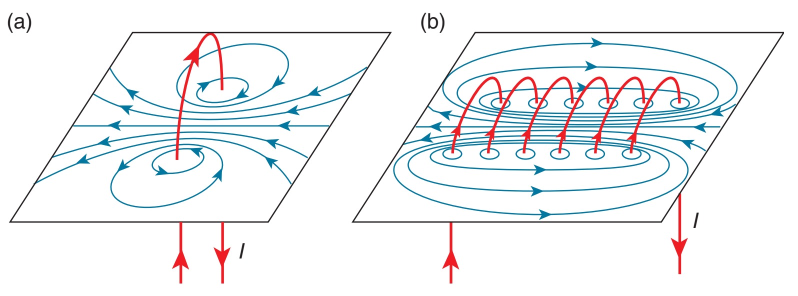


Figure 3.16

(a) The field lines around a single current loop. (b) The field lines become much more concentrated inside a solenoid.

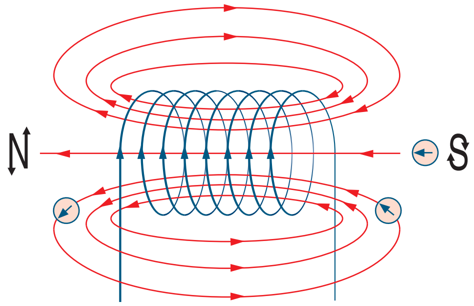


Figure 3.17

This solenoid has an effective 'north' end at the left and a 'south' end at the right. The compass points in the direction of the field lines; that is, towards the south pole of another magnet. A simple way to remember which end of a solenoid corresponds with which pole, is to put arrows on the letters N and S as shown. The arrows indicate the direction of the current as seen from that end.

We saw earlier that a piece of soft iron placed in a magnetic field becomes a temporary magnet. A piece of soft iron can easily be put right inside the solenoid. When the current is turned on, the field of the current induces magnetism in the iron. The field from the iron is now added to that of the current, making the total field very much stronger than that due to the current alone; in fact, it can be 1000 times greater. This arrangement is called an **electromagnet**. Electromagnets in various forms are used in vast numbers in our modern world.

The fact that a relatively small electric current in a coil around a piece of iron can turn the iron into a magnet is a reminder that all magnetism seems to be electrical in its basic nature. This was the obvious question that Ampère asked: If magnetic effects could be produced by electricity alone, what did this imply about magnets? Do you think that, in some way, magnets have tiny electric currents in them?

Physics in action — Magnetic domains

Ferromagnetic substances include iron (and some of its compounds and alloys), nickel, cobalt and gadolinium. Although all atoms have spinning electrons, only a small number of materials are ferromagnetic, and not all pieces of iron are magnets. It turns out that while almost all of the magnetic dipoles do line up with their neighbours, they do it in large groups of atoms, not right throughout the whole piece of metal. The groups contain huge numbers of atoms (around a billion billion, or 10^{18}), but are still rather small in size (of the order of micrometres, 10^{-6} m). These groups of atoms are called *domains*. Normally, the directions of their dipoles are random and so their fields all cancel out.

Each domain has a very strong (but tiny) magnetic field, but if all the domains are oriented differently, the overall effects will cancel out. When the iron is put into a magnetic field, two things can happen: the domains whose magnetism is already pointing in the direction of the field may grow by (in a manner of speaking) taking over nearby groups of atoms, or whole domains may switch their orientation so that they are more in line with the field. The extent to which these things happen determines how strong the magnet will become.

When the external field is removed, the domains may flip back to their random directions (e.g. in soft iron) or they might retain their new directions (e.g. in hard iron). Permanent magnets need to be made out of iron that is as 'hard' as possible; that is, iron in which the domains are very resistant to changes in their orientation once the iron is cool. They are made by heating the iron in a strong field. The heat enables the domains to change more easily.

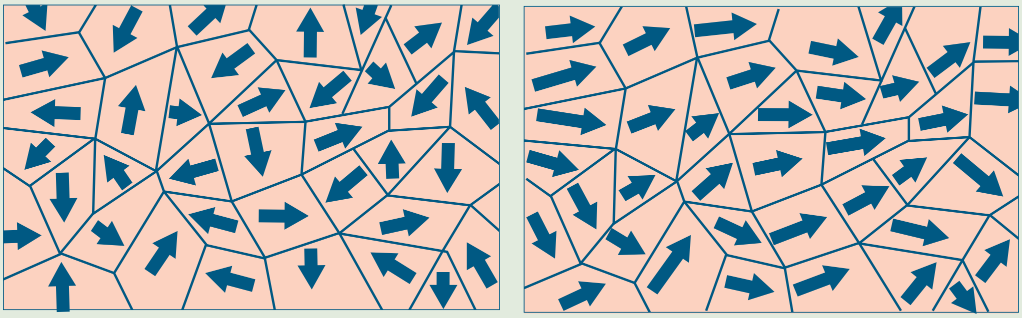


Figure 3.18 When placed in an external field, the domains that are oriented in the right direction grow and the others either shrink or change direction.

Ampère's explanation of the magnetisation of iron was, in fact, remarkably like the modern theory. He felt that all the 'molecules' of iron (we would now say atoms) had their own permanent, closed current loops, making them like little electromagnets. Normally these little 'electromagnets' are oriented in all different directions so that their fields cancel each other. In a permanent magnet, however, many of the little electromagnets are aligned in the same direction, so that their fields add together to make a field big enough to spread out from the magnet into its surroundings. Materials in which this happens are described as **ferromagnetic**. We now know that the atomic electromagnets that arise from a current loop are due to the electrons spinning around in the iron atoms, and physicists have developed a very good picture of the fundamental mechanism of magnetism.

The reason that cutting a magnet in half does not give us separate poles is now clear. A normal magnet is really made up of huge numbers of these tiny electromagnets. Each one of these tiny electromagnets has two 'poles' itself, just as a solenoid does. In fact, at this level it is hardly worth speaking of poles at all; there is a continuous field threading through each little electromagnet. What we have called the poles are really just the two sides of the current loop. No matter how big or small the magnet is, cutting it will not enable us to find a single pole.

Physics file

There are three main types of permanent magnet: alnico, ferrite and rare earth. The major contributor to the magnetism in each of these is iron. Alnico stands for **al**uminium, **ni**ckel and **co**balt, the elements added to iron to make an alloy which has superior magnetic properties to those of normal hard iron.
The most widely used magnets are now made from ferrite, or iron oxide (Fe_2O_3). They are made by grinding the ferrite into a very fine powder, compressing it and then heating it strongly, so they are often referred to as *ceramic ferrites*. Barium or strontium is often added to make a stronger crystal structure. Natural lodestone is a form of ferrite.
In the 1970s it was found that the rare earth element samarium could be alloyed with cobalt to produce a very strong *rare earth magnet*. Newer, less expensive and less fragile rare earth magnets use neodymium, iron and boron.

Practical activity

30 Investigating electromagnetic induction

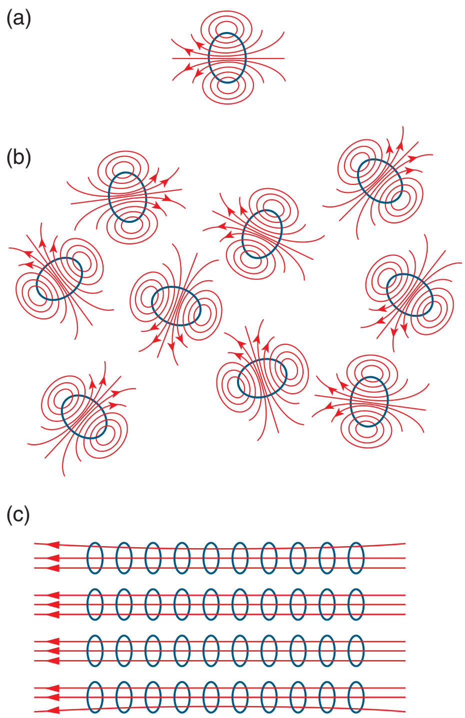


Figure 3.20

(a) A little atomic electromagnet with its magnetic field. (b) Many little electromagnets oriented randomly produce no external field. (c) Many little electromagnets aligned with their fields parallel produce a large external field.

We can also now understand induced magnetism, both in the case of putting a piece of iron near a magnet and in putting a soft iron core in a solenoid. The little electromagnets in the iron become aligned by the influence of the other magnet or current (just as little magnets will line themselves up with each other), and then add their fields to that of the original field. This tends to happen within groups of atoms called *domains*.

Figure 3.19

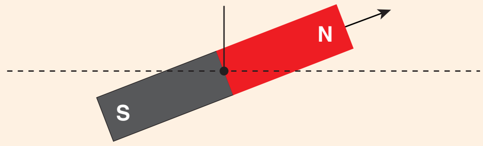
The electromagnet is an extremely useful device. It is used in all sorts of applications, from the tiny, high-speed electromagnets that operate the pins in a dot matrix printer, to the enormous variety of relays (electric switches), and the huge electromagnets used in junk yards to pick up whole cars.

3.1 SUMMARY Magnetic fields

- All magnets are dipolar, i.e. north and south poles always occur together.
- Like magnetic poles repel and unlike magnetic poles attract.
- The Earth has a dipolar magnetic field which acts as a huge bar magnet, with the south end near the geographic north pole.
- The direction of a magnetic field at a particular point is the same as that of the force on the single north pole of a magnet. The direction of the force on the south pole of a magnet is in the direction opposite to the field.
- The magnetic flux density, ***B***, is a vector quantity. Vector addition enables us to add fields and to compare the magnitudes of different fields.
- An electric current produces a magnetic field which is circular around the current. The direction of the field is given by the right-hand grip rule.
- The right-hand grip rule gives the direction of the field produced by a current in a conductor.
- The magnetic field produced by a single loop of current is perpendicular to the loop, and is stronger inside the loop than outside. The magnetic field of a solenoid is much stronger inside than outside.
- A soft iron core in a solenoid can increase the overall strength of the magnetic field up to around 1000 times.
- All magnetism is electrical in nature. At the fundamental level, a magnetic field originates from the electrons orbiting in atoms.
- In ferromagnetic materials, the fields from each atom tend to align so that they add together to produce a greater field. This alignment can be permanent (hard iron) or temporary (soft iron).

3.1 Questions

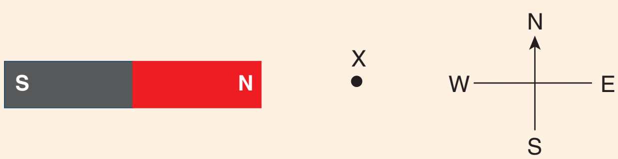
1 The following diagram shows a bar magnet suspended at its midpoint by a light wire. Which of the following is true?



A The magnet is in the northern hemisphere; its N end is pointing towards the geographical south pole.
B The magnet is in the southern hemisphere; its N end is pointing towards the geographical north pole.
C The magnet is in the northern hemisphere; its N end is pointing towards the geographical north pole.
D The magnet is in the southern hemisphere; its N end is pointing towards the geographical south pole.

2 Which of the following statements is true about a magnet attracting a piece of soft iron?
A The soft iron is already a magnet before it interacts with the permanent magnet.
B The permanent magnet induces the soft iron to become a permanent magnet with the opposite pole closest.
C The permanent magnet induces the soft iron to become a temporary magnet with the opposite pole closest.

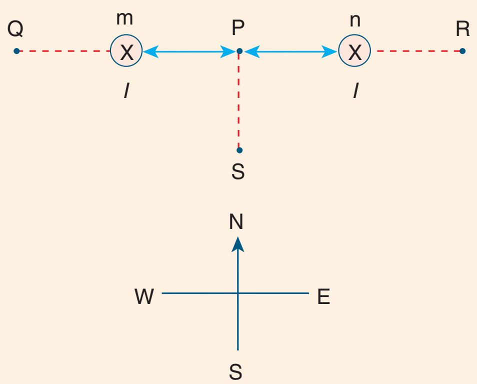
3 In the following diagram, take the strength of the Earth's magnetic field at position X as *B* units. In what direction would a compass point if it was placed at position X and the field strength at X, relative to the Earth's field *B*, due to the magnet shown was:



a $10^{-3}B$? b B? c $10^{3}B$?

The following information applies to questions 4–6.

The following diagram shows a cross-sectional view of two long parallel conductors, m and n, each carrying currents of equal magnitude *I* into the page.

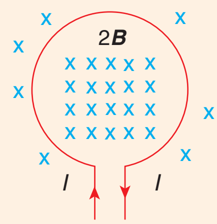


4 a What is the direction of the magnetic field at point P due to conductor m?
b What is the direction of the magnetic field at point P due to conductor n?
c What is the magnitude of the resultant magnetic field at P?

5 At which of the points Q, R and S could the resultant magnetic field due to the currents and the Earth be zero?

6 The direction of the current in conductor n is now reversed.
a What is the direction of the magnetic field at point P due to conductor m?
b What is the direction of the magnetic field at point P due to conductor n?
c What is the direction of the resultant magnetic field at P?

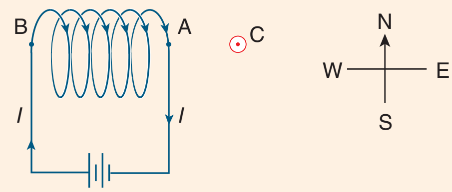
7 The diagram shows a loop carrying a current *I* which produces a field ***B*** in the centre of the loop. It is in a region where there is already a steady field of ***B*** (the same as that due to *I*) directed into the page, so that the resultant field is 2***B***.



What would the magnitude and direction of the resultant field at the centre of the loop become in each of the following cases?
a The current in the loop is switched off.
b The current is doubled.
c The direction of the current is reversed.

The following information applies to questions 8–10.

The diagram shows a current-carrying solenoid located next to a long, straight, current-carrying conductor C (perpendicular to page) that carries a current out of the page.



8 If the direction of the magnetic force on this conductor is north, which point (A or B) represents the north pole of this solenoid?

9 If the direction of the current through the solenoid is reversed, what is the subsequent direction of the magnetic force on C?

10 The direction of the current through the solenoid is returned to its original direction and the direction of the current through C is reversed (into the page). What is the direction of the magnetic force on C?

3.2 Force on current-carrying conductors

Loudspeakers

A speaker has a cone fixed to a coil of wire moving in an external magnetic field. However, a speaker cannot produce sound without being driven by an electronic source. It also requires an enclosure to properly bring the sound to the listener in a controlled manner. Even the most expensive speaker will perform badly if it is incorrectly enclosed.

The choice of what is a really good loudspeaker at least partly comes down to personal preference. It is not always appropriate to select a speaker solely on the basis of frequency response, power handling and distortion, as it is difficult to equate purely scientific measurement with the 'quality' of a speaker's sound.

Moving coil speakers

The principle of the loudspeaker was discovered by Hans Oersted in 1819. It is based on the induced magnetic field that is created when a current moves through a wire. A current moving through a coil creates a magnetic field that interacts with the magnetic field from a permanent magnet. If the current moves in one direction, it might make the coil move to the left; reversing the current will cause it to move to the right. Figure 3.21 shows how an alternating current makes the coil around the fixed magnet move backwards and forwards, creating compressions and rarefactions in the air in front of the cone. In the case of a simple sine wave, the applied current will reverse direction from positive to negative and vice versa, passing through 0 V in between. The resulting sound is a single clear tone.

The larger the current that moves through the wire, the larger the induced magnetic field will be, and hence the greater the force of magnetic attraction or repulsion that will be exerted by the permanent magnet. The speaker cone will move in and out through a greater amplitude, transferring more energy to the surrounding air molecules and thus creating a louder sound. You can see this happening by watching the movement of an open speaker as the electrical signals from an amplifier are passed through it.

In practice, the simple bar magnet of Figure 3.21 is replaced by a ring magnet. The coil then vibrates in a strong radial field. The diaphragm and cone are generally composed of paper or stiff plastic. Flexible edge suspensions, or springs, surround the outer edge and the central diaphragm. These springs resist the force of the speaker's movement and provide a restoring force to the cone: they return the cone to a central rest position after it has been driven forwards or backwards by an electrical signal.

Today almost all loudspeakers are moving coil speakers. They come in a wide variety of sizes, roughly matching the range of frequencies they have been designed to best produce: 'woofers' for frequencies from 30–500 Hz, mid-range loudspeakers for 500–4000 Hz, and the aptly named small 'tweeters' for the high frequencies from 4–20 kHz.

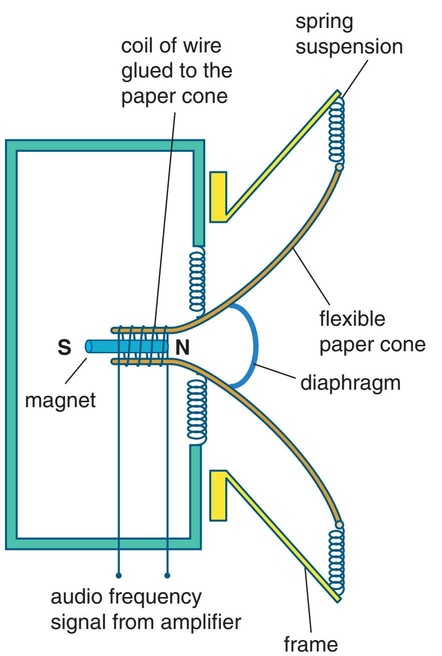


Figure 3.21
The principle on which a dynamic, or moving coil, loudspeaker operates. The changing current through the coil moves the magnet to and fro, making a paper or plastic cone vibrate, which in turn causes the air in front of it to vibrate. This can be clearly seen when the grill is removed from the front of a speaker, leaving it open to the air. A candle flame placed in front of the speaker will move back and forth.

Figure 3.22
A modern hi-fi loudspeaker system will often have a range of speakers, each best suited to producing a particular range of frequencies. In general, the larger the speaker, the lower the frequencies it will reproduce.

The effect of a magnetic field on an electric current

The fact that an electric current responds to an external magnetic field can most easily be seen by letting a wire hang freely near a magnet. When a current drifts in the wire, the wire moves—there is a magnetic force acting on it. Further investigation reveals that the force is a maximum when the current is *perpendicular* to the field, and drops to zero if the current and field are parallel. The direction of the force is perpendicular to *both* the current and the field.

A simple **right-hand force rule** helps remember these relative directions. Orient your right hand so that your thumb points in the direction of the conventional current (as for the previous right-hand grip rule, section 3.1) and your fingers, held straight this time, are in the direction of the external field (**fi**ngers = **fi**eld). The force is then in the direction we would normally push: out from our palm.

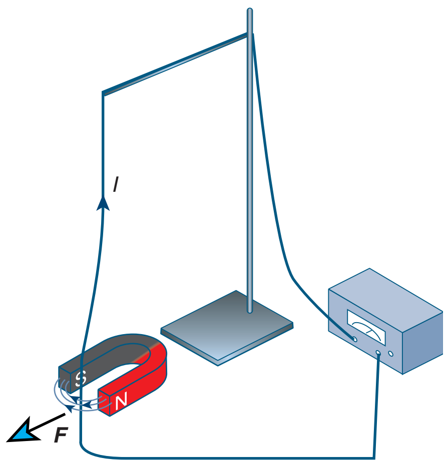


Figure 3.23
When current flows in the flexible wire between the poles of a strong U magnet, the wire experiences a force at right angles to the current and to the field.

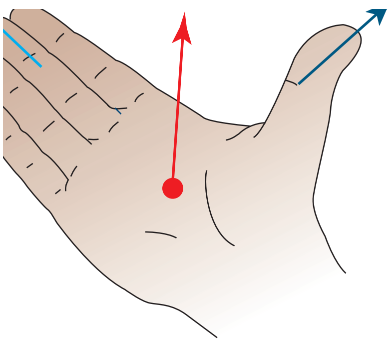


Figure 3.24
The right-hand force (or palm) rule gives the direction of the force, ***F***, on a current, *I*, in a magnetic field, ***B***. The thumb represents the direction of the current, the fingers represent the direction of the field, and the direction of the force is out from the palm of the hand, as you would push.

It is very important not to get the two right-hand rules mixed up, as they apply to quite different situations. The right-hand grip rule (Figure 3.12, section 3.1) tells us the direction of the circular magnetic field that occurs around a current-carrying conductor. *This field is created by the current*. The right-hand force rule (sometimes called the right-hand palm rule or just the right-hand rule) tells us the direction of the

Physics file

Loudspeakers are capable of storing energy. This happens whenever the cone is displaced from its rest position. After the driving force is removed, the springs try to return the cone to its rest position, resulting in speaker oscillation. For acoustic reasons, the vibration of the speaker cone must be limited or damped or the sound will become distorted.
One way to dampen the cone's motion is to apply a counter EMF by shorting the coil terminals. Self-induction causes the coil to generate a counterforce which opposes the cone's movement. This effect is referred to as a 'back EMF'. How well a speaker succeeds in reproducing real sounds will depend, at least to a degree, on how well the damping can be applied by the amplifier.

Practical activity

29 Direction of induced current in a wire

force on a current in a magnetic field (Figure 3.25). *This force is the result of placing a current in a magnetic field.* Do not confuse it with the field created by the current itself.

Ampère showed that there is also a magnetic force between two parallel electric currents. We can regard one as creating a magnetic field and the other as responding to that field. You can use the two right-hand rules in turn to show that if the currents flow in the same direction, the force is attractive, and that if they flow in opposite directions, the force is repulsive (Figure 3.25). The cause of the magnetic force in the wire is the interaction between the induced magnetic field created by the current in the conductor and the external magnetic field. The effect is exactly the same as the attraction or repulsion that we see between two permanent magnets.

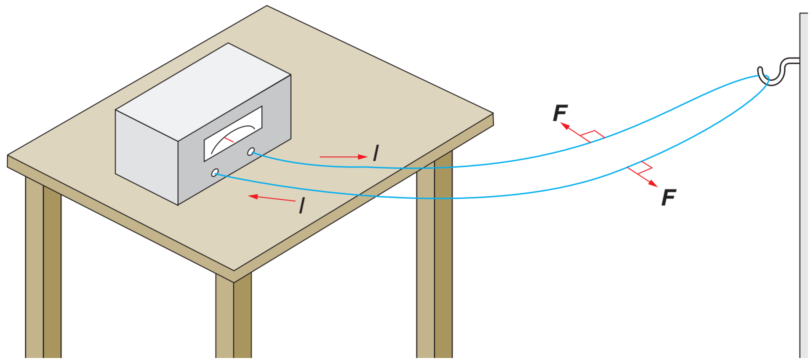


Figure 3.25
Two wires with a current flowing in opposite directions experience a repulsive force. You can repeat this experiment with a length of flexible wire connected to a power supply and looped over a hook on the wall. Have the wires about 2 cm apart and at least 2 metres long, and use the maximum current that is safely available from the power supply. When you turn the power on and off the wires should swing out as a result of the repulsive force.

Physics file

You might realise that earlier we effectively defined the relative strength of a magnetic field in terms of the torque forces on a compass needle. That this is equivalent to the force on an electric current will become clear as we discuss the real nature of magnets.

Calculating the magnetic force

The force on a wire carrying a current in a magnetic field with a perpendicular magnetic flux density ($B_{\perp}$) clearly depends on the amount of current flowing (I) and the length of wire in the field (l). Simple experiments will soon confirm that indeed the force is directly proportional to each of these, i.e. $F_B \propto Il$.

This is not surprising. If we had two wires, each with the same current, and tied them together in the magnetic field, we might well expect that the force on the pair would be double the force on each one separately. We could see this experiment as either doubling the current or doubling the length of the wire in the field. In either case, we are not surprised to find twice the force.

What do we mean by the magnetic field $\boldsymbol{B}$? In discussing the gravitational field $\boldsymbol{g}$ and the electric field $\boldsymbol{E}$, the meaning of a field was clear; it was the force per unit of mass or unit of charge, respectively, and the units were newtons per kilogram and newtons per coulomb. But what does the magnetic force act upon? The answer is that it acts on a *length of current*. The perpendicular magnetic flux density is then the force per unit length of unit current. That is:

$$B_{\perp} = \frac{F}{I\,l}$$

We can write this relationship as $F = IlB_{\perp}$, and so we see that the constant in our earlier proportionality is B, the strength of the magnetic field.

$$F = IlB_{\perp}$$

where F is the force on a wire carrying a current in a magnetic field (N), I is the conventional current in the wire (A), l is the length of wire in the field (m), and $B_{\perp}$ is the perpendicular magnetic flux density (T).

Earlier when we talked of the strength of a magnetic field, we put off the question of the units for magnetic field. Now we can see that there is a natural way to measure the strength of a field in terms of the force F (in newtons) on a current I (in amperes) in a field of length l (in metres). We say that the unit for magnetic flux density will be the field that produces a force of 1 newton on a 1-metre length of wire with a current of 1 ampere. The unit for magnetic flux density then becomes the newton per amp metre ($N\ A^{-1}\ m^{-1}$). This unit was given the name *tesla* (T) in honour of Nikola Tesla. A field of 1 **tesla** is a very strong field. For this reason, a number of smaller units (especially the millitesla, 10^{-3} T, and microtesla, 10^{-6} T) are in common use.

1 ***tesla*** = 1 newton per amp metre ($1\ T = 1\ N\ A^{-1}\ m^{-1}$)

The Earth's magnetic field at the surface is around 5×10^{-5} T. This is equivalent to 0.05 mT, or 50 μT. You are likely to come across both these units, as well as the gauss! (See Physics file this page.) Due to the shape of the Earth and the shape of the Earth's magnetic field, the angle at which the field emerges and enters the ground varies depending on the latitude. The angle up or down from horizontal is known as magnetic inclination; it is positive when the field points down as it does in the northern hemisphere and negative when the field is upwards as in the southern hemisphere. Directly above the south magnetic pole, the Earth's magnetic field comes out of the ground at an inclination of –90.0°, while at the north magnetic pole it enters the ground at a dip angle of +90.0°. At the equator, the magnetic field is approximately horizontal, while in Perth, in 2005, the Earth's field had a magnitude of 5.3406×10^{-5} T, decreasing by 33 nT per year, and a magnetic inclination of –55.899°, which is increasing by 0.030° per year.

There is often some confusion around the name and units of B. The name given to B is the *magnetic flux density*, but you may also see the terms: *magnetic induction* or the *strength of the magnetic field*. We have been careful to avoid the term 'magnetic field strength', which is actually the name of a slightly different quantity in more advanced physics.

Physics file

An older unit for B, the gauss, based on the cgs system of units (centimetre/gram/second) is also commonly used: 1 gauss = 10^{-4} tesla or 100 μT. The Earth's magnetic field has an intensity of around 0.5 gauss, or 500 mG.

Physics file

Nikola Tesla (1856–1943) was the first person to advocate the use of alternating current (AC) generators for use in town power supply systems. He was also a prolific inventor of electrical machines of all sorts, including the Tesla coil, a source of high-frequency, high-voltage electricity.

Physics file

The Australian Government's Geoscience Australia division has a website where you can enter your location and get the magnetic details of your region. Search the Internet with these key words to locate the website: Australian geomagnetic reference field values.

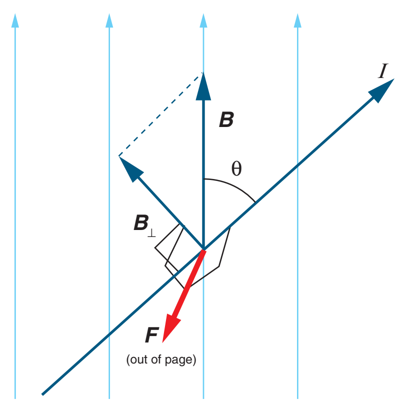


Figure 3.26
The magnitude of the force on a current depends on the component of the field at right angles to the current. The component of the field at right angles to the current in this situation is $B_{\perp} = B \sin\theta$.

Physics file

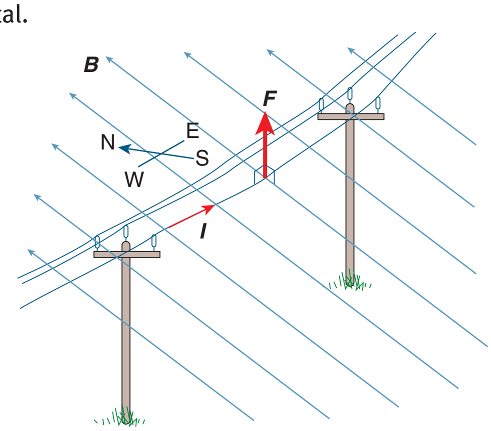
The three quantities force, length and field are all vectors (***F***, ***l*** and ***B***). To express the relationship between them fully we need what is called 'vector cross multiplication'. If we write $\boldsymbol{F} = I\boldsymbol{l} \times \boldsymbol{B}$, this is taken to mean that the magnitude of vector ***F*** is the product $IlB\sin\theta$, where θ is the angle between ***l*** and ***B***, and the direction is at right angles to both ***l*** and ***B*** and in the sense given by the right-hand rule.

table 3.1 Some typical magnetic fields

Magnetic field	Field (T)
The surface of a neutron star	10^8
Largest artificially produced pulsed fields	10^3
Very strong electromagnets and 'supermagnets'	1–20
Magnetic storms on the Sun	1
Alnico and ferrite magnets	$10^{-2} - 1$
The Earth's surface	5×10^{-5}
Interstellar space	10^{-10}
Smallest value achieved in a magnetically shielded room	10^{-14}

Worked Example 3.2A

Determine the magnetic force per metre, due to the Earth's magnetic field, on a suspended power line running east–west and carrying a current of 100.0 A, as shown in the diagram. Assume the magnetic field is at a magnetic inclination of 55.9° up from horizontal.



Solution

As the line is running east–west, the current and magnetic field are at right angles. If we assume the Earth's magnetic flux density is 5.34×10^{-5} T, the magnitude of the force on 1.00 m of this power line is given by:

$B_\perp = 5.34 \times 10^{-5}$ T

$I = 100.0$ A

$l = 1.00$ m

$F_B = IlB_\perp$

$= (100.0)(1.00)(5.34 \times 10^{-5})$

$= 5.34 \times 10^{-3}$ N per metre at 90° to ***l***

The direction of the force will be 55.9° up from horizontal to the south. In practice, the current in a power line is normally AC and so the force would alternate in direction from 55.9° up from horizontal to the south to 55.9° down from horizontal to the north.

We have seen that an electric current that is placed in a magnetic field that is at right angles to the current experiences a force that is at right angles to both the current and the field. If the field is not at right angles, the force is correspondingly less. In fact, if the field and current are parallel, there is no force at all. The general formula to describe the relationship between $\boldsymbol{B}_\perp$ and ***B*** is:

$$\boldsymbol{B}_\perp = \boldsymbol{B} \times \sin\theta$$

where $\boldsymbol{B}_\perp$ is the magnetic flux density at right angles to the length of the conductor (T), ***B*** is the magnetic flux density at the location of the conductor (T) and θ is the angle between ***I*** and ***B***.

✓ Worked Example 3.2B

If the 100.0 A power line in Worked Example 3.2A runs north–south instead of east–west, and the magnetic inclination of the Earth's field is 55.9° up from horizontal, what would be the force on the line per metre of line?

Solution

This time the field and current are not at right angles and so we need to determine $\boldsymbol{B}_{\perp}$. The magnetic inclination is the angle of the field up from the horizontal and so it is also the angle between the current and the field.

$\boldsymbol{B}_{\perp} = 5.34 \times 10^{-5}$ T
$\theta = 55.9°$

$$\boldsymbol{B}_{\perp} = \boldsymbol{B} \times \sin\theta = (5.34 \times 10^{-5})(\sin 55.9°) = 4.42 \times 10^{-5} \text{ T}$$

Now we can calculate the force on the wire

$\boldsymbol{B}_{\perp} = 4.42 \times 10^{-5}$ T
$I = 100.0$ A
$l = 1.00$ m

$$F_B = I l \boldsymbol{B}_{\perp} = (100.0)(1.00)(4.42 \times 10^{-5}) = 4.42 \times 10^{-3} \text{ N per metre at } 90° \text{ to } l$$

The right-hand rule enables us to find the direction of the force as east if the current is flowing north, or to the west if the current is flowing south.

3.2 SUMMARY Force on current-carrying conductors

- The magnitude of the force on a length of wire l carrying a current I in a magnetic field $\boldsymbol{B}$ is given by:
  $$F = I l \boldsymbol{B}_{\perp}$$
- The component of the magnetic field that is at 90° to the length of wire $\boldsymbol{B}_{\perp}$ is $\boldsymbol{B}_{\perp} = \boldsymbol{B} \times \sin\theta$ where θ is the angle between the current I and field $\boldsymbol{B}$.
- The unit for magnetic field, $\boldsymbol{B}$, is the tesla. A current of 1 A in a field of 1 T will experience a force of 1 N on each metre of its length in the field. A field of 1 T is a very strong field.
- The direction of the force is given by the right-hand palm rule and is perpendicular to both the field and the current.

3.2 Questions

The following information applies to questions 1–3.

An east–west power line of length 100.0 m is suspended between two towers. Assume that the strength of the magnetic field of the Earth in this region = 5.34×10^{-5} T.

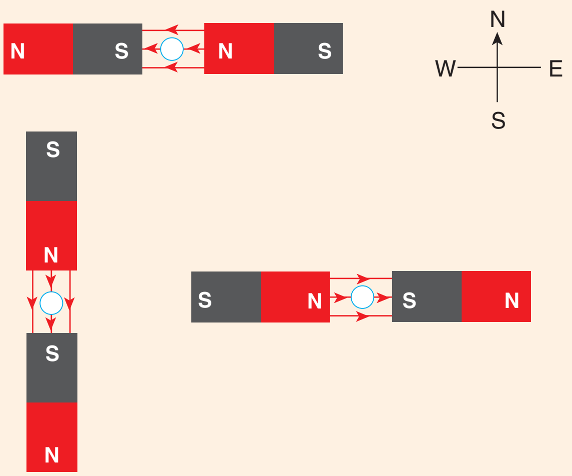
1 Calculate the magnetic force on this power line if it carries a current of:
   a 80.0 A from west to east
   b 50.0 A from east to west.

2 Assuming that $g = 9.80$ N kg^{-1} and that the total mass of the power line is 50.0 kg, calculate the value of the ratio of the weight to the magnetic force on the power line when carrying a current of 100.0 A.

3 Over time, the ground underneath the eastern tower subsides, so that the power line is lower at that tower. Assuming that all other factors are the same, would the magnitude of the magnetic force on the power line in this new situation be:
   A greater than before?
   B the same as before?
   C less than before?

The following information applies to questions 4–6.

The diagram below depicts cross-sectional views of three long, straight, current-carrying conductors, each located between the poles of a permanent magnet. The magnetic field, $\boldsymbol{B}$, of these magnets, and the currents, I, are mutually perpendicular in all cases.



4 For diagram (a), calculate the magnitude and direction of the magnetic force on a 5.00 cm section of conductor with:

a $I = 2.00$ A into the page, $B = 2.00 \times 10^{-3}$ T

b $I = 1.00$ A out of the page, $B = 2.00 \times 10^{-3}$ T

5 For diagram (b), calculate the magnitude and direction of the magnetic force on a 1.00 mm section of conductor for:

a $I = 3.00$ A into the page, $B = 0.500$ T

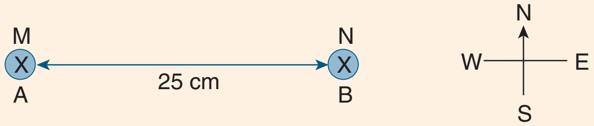
b $I = 3.00$ A out of the page, $B = 1.00$ T

6 For diagram (c), calculate the magnitude and direction of the current that would result in a magnetic force on a 2.00 mm section of conductor of:

a 8.00×10^{-3} N south, for $B = 0.100$ T

b 2.00×10^{-2} N north for $B = 0.500$ T

7 The following diagram shows two parallel, current-carrying conductors, M and N, located at points A and B, each carrying currents of 5.00 A into the page.



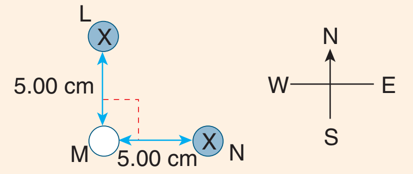
The magnetic field at point B due to M is equal to 4.00 mT south. Calculate the magnitude and direction of the magnetic force per metre of conductor on:

a N

b M

The following information applies to questions 8 and 9.

Three current-carrying conductors L, M and N, each with their axes perpendicular to the plane of the paper, are located as shown in the following diagram. The resultant magnetic field at the position of M is equal to 1.00×10^{-3} T north-west.



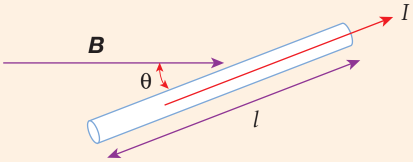
8 Calculate the magnitude and direction of the magnetic force per metre on M for a current of:

a 2.0 A flowing through M into the page

b 1.0 A flowing through M out of the page.

9 If the current flowing in conductor L is switched off, what would be the direction of the resultant magnetic field at the point where M is located?

10 The following diagram shows a conductor of length *l*, carrying a current of magnitude *I*, in an external magnetic field of magnitude *B*.



The angle between $\boldsymbol{B}$ and $\boldsymbol{I}$ is θ. Calculate the magnitude of the magnetic force on this conductor for:

a $B = 1.00$ mT $I = 1.00$ mA $l = 2.00$ mm $\theta = 90.0°$

b $B = 1.00$ T $I = 100$ A $l = 5.00$ cm $\theta = 0.00°$

c $B = 0.100$ T $I = 5.00$ A $l = 1.00$ mm $\theta = 30.0°$

3.3 Electric motors

Physicists have always been interested in the relationship between electricity and magnetism because they want to understand the basic workings of the Universe. For the world at large, however, this understanding provided a more practical form of excitement. It enabled the generation and use of electricity on a large scale. One of the most obvious applications of the understanding of electromagnetism gained last century is the electric motor.

The operating principle of all electric motors is simple. An electric current in a magnetic field experiences a force: $\boldsymbol{F} = I\boldsymbol{l}\boldsymbol{B}_{\perp}$. All direct-current motors work by the action of this force on a wire loop containing an electric current. Normally, several coils of many turns of wire are used. They are placed in a magnetic field provided by either a permanent magnet or an electromagnet.

Consider a wire loop ABCD carrying a current, $\boldsymbol{I}$, initially horizontal in a magnetic field, $\boldsymbol{B}$ (Figure 3.28). Current is flowing clockwise around this coil in the direction A → B → C → D. Because sides AD and BC are parallel to the magnetic field, there is no force on them. Sides AB and CD are perpendicular to the field, so there will be a downward force on AB and an upward force on CD (Figure 3.28a). These two forces will produce a force couple, or torque, on the coil which will tend to rotate it anticlockwise. If the coil is free to turn, a little later it will be in the position shown in Figure 3.28b.

Now all four sides will experience some magnetic force, but the forces on the two sides AD and BC are in line but in opposite directions. They simply tend to stretch the coil. The forces on AB and CD are still the same, but the torque is less because the line of action of the force is closer to the axis of rotation. Soon the coil will be perpendicular to the field (Figure 3.28c), and the forces will try to keep the coil in this position. There is no torque on the coil, but any displacement from the vertical position will result in a torque that will cause the coil to rotate back to this position.

To make an electric motor out of this device, the direction of the current needs to be reversed at this point. All the forces are then reversed, and the situation is as shown in Figure 3.28d. Provided the coil has a little momentum it will pass the vertical position, and the now reversed forces on AB and CD will again create an anticlockwise torque, so the coil will continue to rotate. After every half-turn, the direction of the current must be switched to maintain the anticlockwise torque. To do this, the current is fed to the coil via a **commutator**—a cylinder of copper on which conducting brushes (usually carbon blocks) rub. It is made in two halves, as in Figure 3.29. Each half is connected to one end of the coil of wire. As the coil rotates, the commutator reverses the current at just the right moment.

Interactive tutorial

DC motors

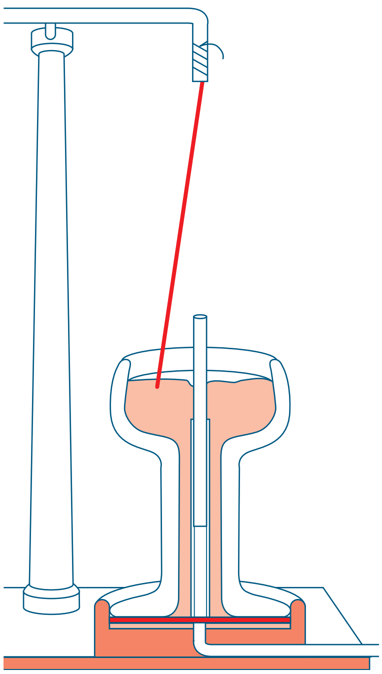


Figure 3.27
In 1821, Faraday built what could be called the first electric motor. A magnet was mounted vertically in a pool of mercury. A wire carrying a current hung from a support above. (The mercury provided a path for the current.) The magnetic field of the magnet spreads outwards from the top of the magnet and so there is a component perpendicular to the wire. This produces a horizontal force on the wire which will keep it rotating around the magnet. Use the right-hand rule to convince yourself that if the current flows down and the magnetic field points up and out, the wire will rotate clockwise.

Practical activity

32 Current from an electric motor

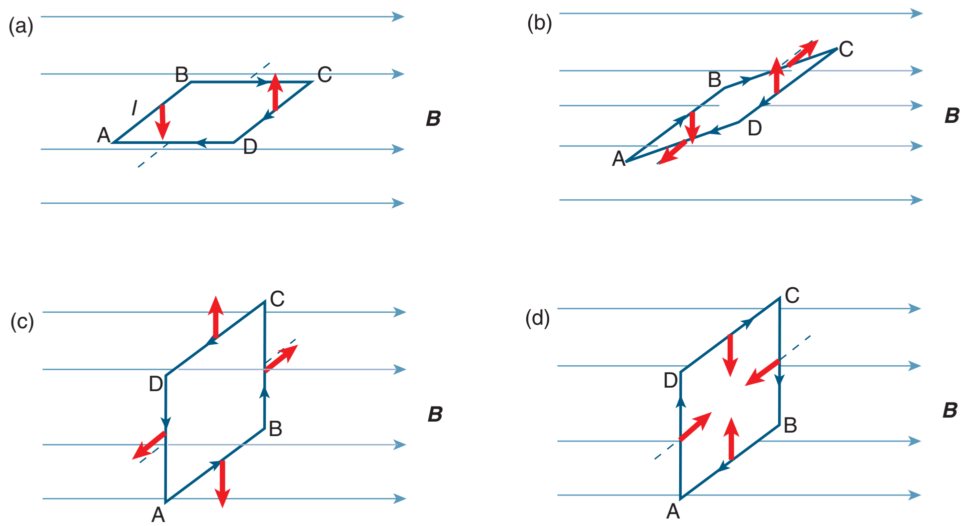


Figure 3.28
Forces acting on a current-carrying wire loop in a magnetic field

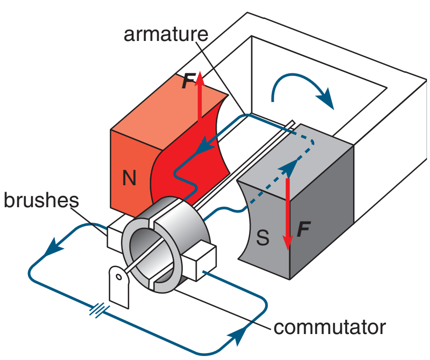


Figure 3.29
The commutator switches the direction of the current every half-turn in order to produce a torque which continues to rotate the coil in the same direction.

This type of motor would be rather jerky, receiving maximum torque only twice every turn, so practical motors are usually made with many coils, all spaced at an angle to each other, and the commutator is arranged to feed current to the coil that is in the best position for maximum torque. The coils are wound on a soft iron core to intensify the magnetic field through them. The whole arrangement of core and coils is called an *armature*. Permanent magnets are often used to provide the magnetic field in small motors, but in larger motors, electromagnets are used because they can produce larger and stronger fields. Because these magnets are stationary, as distinct from the rotating armature, they are often referred to as the *stator*.

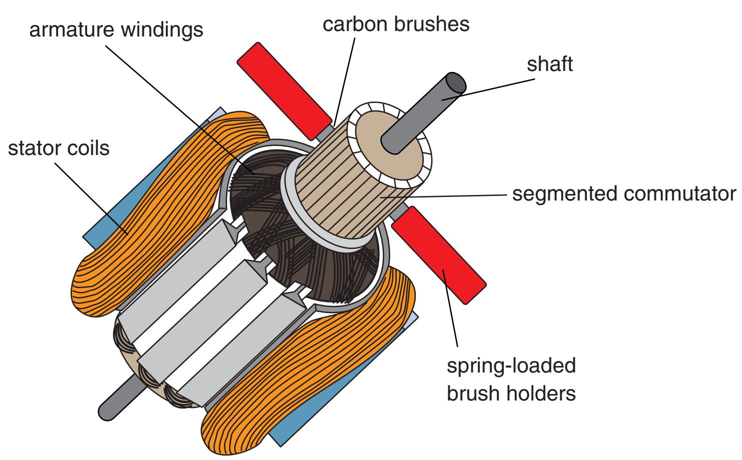


Figure 3.30
A typical universal electric motor, showing the main components. Some motors would have additional stator coils. The commutator feeds current to the armature coils in the position where most torque will be experienced. This type of motor will operate on either DC or AC current.

✓ Worked Example 3.3A

The coil shown in Figure 3.28a is 4.00 cm square and consists of 20 turns of wire and carries a current of 5.00 A in a field of 0.800 T.

a What are the forces on each side?

b What is the torque on the coil?

Solution

a There are 20 turns in the coil so there are 20 lengths of wire on each side. The forces on sides AD and BC are zero because they are parallel to the field. The magnitude of the force on the other two sides is:

$N = 20$ turns

$I = 5.00$ A

$l = 0.0400$ m

$B_{\perp} = 0.800$ T

$$F = NIlB_{\perp}$$
$$= (20)(5.00)(0.0400)(0.800)$$
$$= 3.20 \text{ N}$$

The force on AB is down and on CD it is up.

b The torque on one arm of the coil is:

$F = 3.20$ N

$r_{\perp} = 0.0200$ m

$$\tau = Fr_{\perp}$$
$$= (3.20)(0.0200)$$
$$= 6.40 \times 10^{-2} \text{ N m}$$

The total torque will be twice this as both arms contribute the same torque, so:

$$\tau_{total} = 2 \times \tau$$
$$= 2(6.40 \times 10^{-2})$$
$$= 1.28 \times 10^{-1} \text{N m}$$

The direction of the torque is anticlockwise when viewed from the front.

Generally speaking, a higher torque in an electric motor is preferable. This is achieved by the use of a strong field, a large number of turns of wire, a high current and a large area of coil. All this adds to the cost, so the design of a motor has to be carefully considered in the light of its potential use.

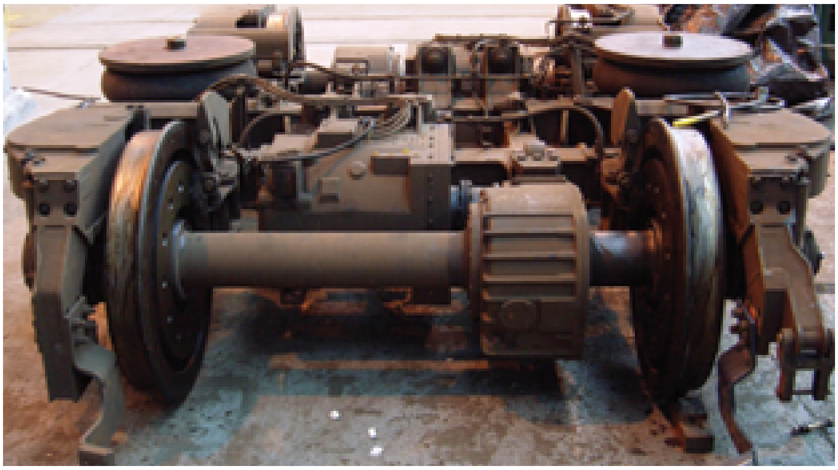


Figure 3.33
This is the power bogie of a modern electric train like those on the Perth rail system. The three-phase AC motor (centre left) drives the wheels via the gearbox (centre right). A complex electronic system converts the 1500 V DC from the overhead lines to variable frequency, variable voltage AC current as required by the speed and load conditions of the train. The motor acts as a generator in order to slow the train and feeds the current back into the line.

Physics file

A *torque* is the turning effect of a force. For example, you apply a torque to a nut with a spanner. For maximum effect the force you apply should be at right angles to the spanner. The larger the force and the longer the spanner, the greater the torque. The torque, τ, is defined by $\tau = Fr_{\perp}$, where $r_{\perp}$ is the perpendicular distance between the line of action of the force and the pivot point. This is shown in Figure 3.31.

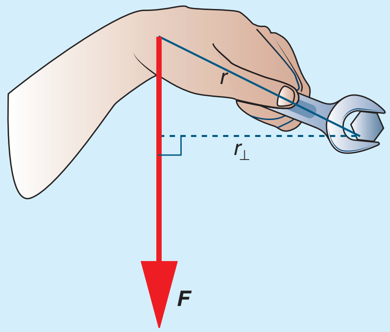


Figure 3.31

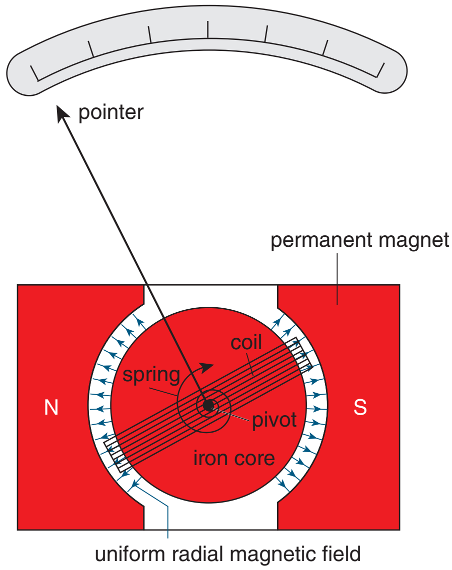


Figure 3.32
The movement of a galvanometer. The shaped pole pieces and the iron core in the centre intensify the field as well as making it more uniform. This is important if the meter is to be accurate.

Physics file

Older 'analogue' voltmeters and ammeters also use the principle of a coil in a magnetic field. In this case, there is no need for a commutator; the coil, with needle attached, turns against a spiral spring. The greater the current through the coil, the further it turns against the spring. Only a very small current is needed to deflect the coil through its full range. This basic apparatus is called a *galvanometer*.
In a voltmeter a high resistance is placed in series with the coil of the galvanometer, and the small current through the coil is a measure of the potential difference across the terminals. An ammeter has a low resistance in parallel with the coil. Only a small proportion of the current goes through the coil, but this is an indication of the total current flowing.

The motors used in food mixers, electric drills and so on are known as universal motors because they can be used on either AC or DC power. The key to using a motor of the type we have been discussing on alternating current is to use an electromagnet for the stator and a large number of coils in the armature. If the stator and armature are both fed with alternating current, the direction of the force on the coil remains constant even though the field reverses 50 times each second, because the coil current reverses as well. The large number of coils in the armature is needed to ensure that there are always loops in the best position to produce maximum torque, at the peak of the current cycle.

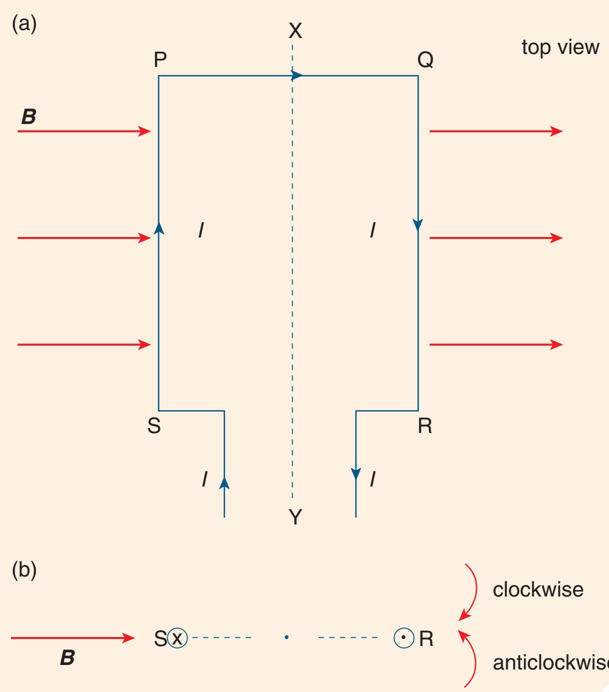
3.3 SUMMARY Electric motors

- An electric motor relies on the magnetic force on a current-carrying loop in a magnetic field:
$$F = IlB_{\perp}$$
- If there is more than one turn on the coil (N) then the force on each side is found by:
$$F = NIlB_{\perp}$$
- There is a torque on the loop whenever its plane is not perpendicular to the field.
- The loop keeps rotating because the direction of current, and hence torque, is reversed each half turn by the commutator.
- The armature of a practical motor consists of many loops that are fed current by the commutator when they are in the position of maximum torque.
- The torque depends on the strength of the field, the current, the number of turns and the area of the coils.

3.3 Questions

The following information applies to questions 1–5.

Diagram (a) at right depicts a top view of a current-carrying loop in an external magnetic field ***B***. Diagram (b) is the corresponding cross-sectional view as seen from Y. The following data applies to this diagram: ***B*** = 0.100 T, PQ = 2.00 cm, PS = QR = 5.00 cm, *I* = 2.00 A.



1 Calculate the magnitude and direction of the magnetic force acting on side PS.

2 Calculate the magnitude and direction of the magnetic force acting on side QR.

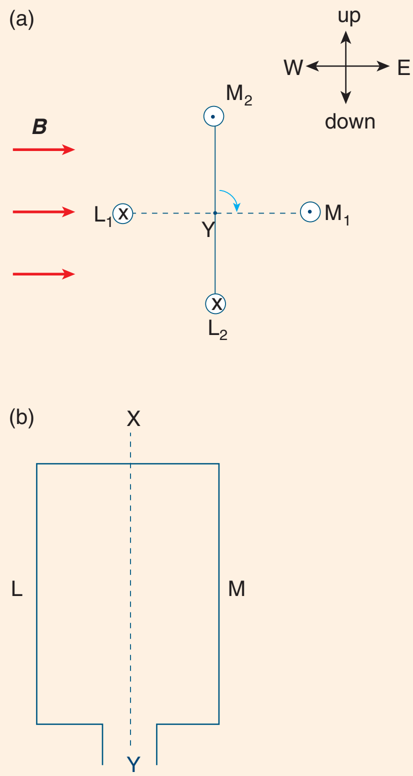
3 What is the magnitude of the force on side PQ?

4 This loop is free to rotate about an axis through XY. In what direction would this loop rotate (as seen from Y)?

5 Which of the following does not affect the magnitude of the torque acting on this loop?
- **A** the dimensions of the loop
- **B** the magnetic field strength
- **C** the magnitude of the current through the loop
- **D** the direction of the current through the loop

The following information applies to questions 6–9.

One of the coils in the electric motor in a hybrid car is illustrated in diagram (a) at right and shows a cross-sectional view of the sides L and M of a current-carrying loop, located in an external magnetic field of magnitude *B* directed east. The corresponding top view of this loop is shown in (b). Note the current directions. This loop is free to rotate about an axis through XY. With reference to diagram (a):



6 When LM is aligned horizontally (L_1–M_1), what is the direction of the magnetic force on side:
- **a** L?
- **b** M?

7 In what direction will the loop rotate?

8 When LM is aligned vertically (L_2–M_2) what is the:
- **a** direction of the magnetic force on side L?
- **b** direction of the magnetic force on side M?
- **c** magnitude of the torque acting on the loop?

9 When LM is aligned vertically, which one of the following actions will result in a torque acting on the loop that will keep it rotating in an anticlockwise direction? (Assume it still has some momentum when it reaches the vertical position.)
- **A** Increase the current through the loop.
- **B** Increase the magnetic field strength.
- **C** Reverse the direction of the current through the loop.

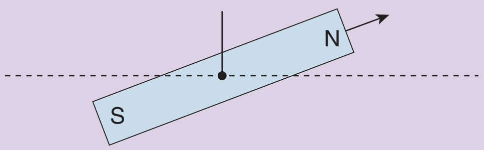
10 Briefly explain the function of the commutator in an electric motor.

Chapter 3 Review

Questions indicated by * relate to material on the CD.

1 If m carries a current into the page and n carries a current out of the page, at which of the following points could a third conductor experience zero magnetic force?
- **A** a point between m and n
- **B** a point west of m
- **C** a point east of n

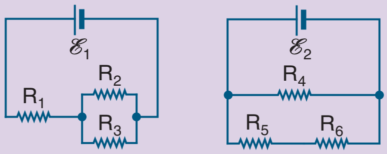
2 The following diagram shows a bar magnet suspended by a light wire. Which of the following is true?



- **A** The magnet is in the northern hemisphere and is pointing towards the geographical south pole.
- **B** The magnet is in the southern hemisphere and is pointing towards the geographical north pole.
- **C** The magnet is in the northern hemisphere and is pointing towards the geographical north pole.

3 Repeatedly cutting a magnet in half always produces magnets with two opposite poles. From this information, which of the following can be deduced in relation to the poles of the magnet?
- **A** The soft iron is already a magnet before it interacts with the permanent magnet.
- **B** The permanent magnet induces the soft iron to become a permanent magnet with the opposite pole closest.
- **C** The permanent magnet induces the soft iron to become a temporary magnet with the opposite pole closest.

***4** In the following two circuits the cells and resistors are identical. Assume the cells are ideal (i.e. no internal resistance).



- **a** Which resistor(s) has the highest current flowing through it?
- **b** Which resistor(s) has the lowest current flowing through it?
- **c** Which resistor(s) has the highest power dissipated through it?
- **d** Which cell is supplying the largest current?

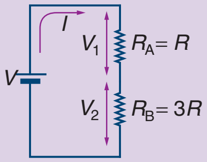
***5** You have four 4.00 Ω resistors. How would you arrange the four resistors to give a total effective resistance of:

a 16.0 Ω
b 10.0 Ω
c 1.0 Ω
d 5.33 Ω

***6** If the resistors in question 5 were connected to a 12.0 V battery, calculate:

a the total current flowing into the circuit for each arrangement.
b the current in each 4.00 Ω resistor in the arrangement.

The following information applies to questions 7 and 8.
The circuit shows two resistors connected in series across a cell with a terminal potential difference of ΔV. The resistance of R_B is three times that of R_A.

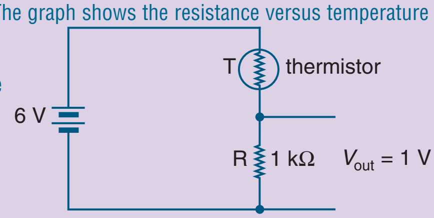


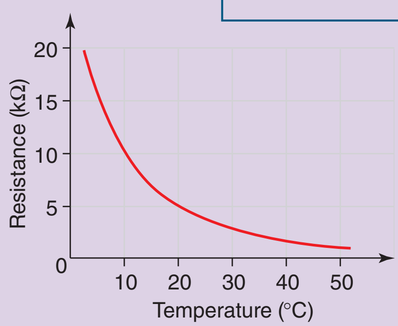
***7** Show mathematically that one-quarter of the cell's voltage is dropped across R_A.

***8** If $V = 12.0$ V and $I = 200.0$ mA, determine the:

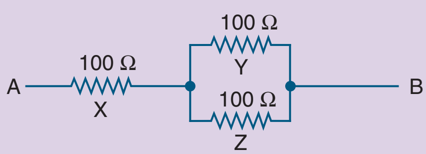
a resistance values of R_A and R_B.
b power dissipated in R_B.

***9** A thermistor is a semiconductor device whose resistance depends on the temperature. The graph shows the resistance versus temperature of the thermistor. Determine the temperature of the thermistor in the circuit if V_{out} is 1 V.



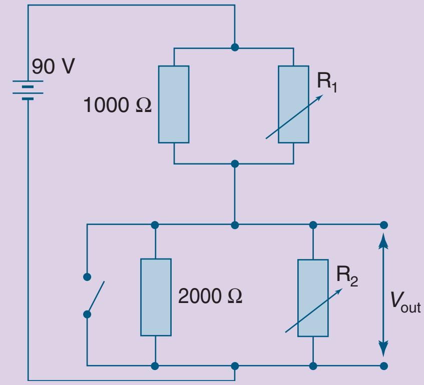


***10** Three 100 Ω resistors are connected as shown. The maximum power that can safely be dissipated in any one resistor is 25 W.



a What is the maximum potential difference that can be applied between points A and B?
b What is the maximum power that can be dissipated in this circuit?

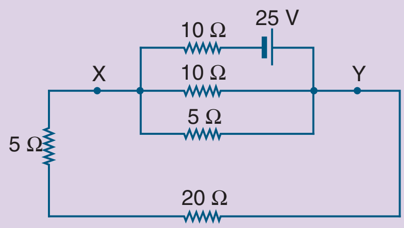
***11** The circuit below combines variable resistors R_1 and R_2 with fixed resistors to make a complex voltage divider. Copy and complete the table by determining the output voltage for each row.



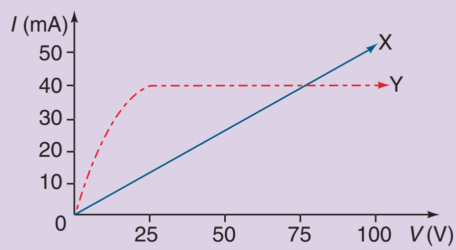
R_1 (Ω)	R_2 (Ω)	Switch	V_{out} (V)
1000	2000	Open	
2000	4000	Open	
4000	2000	Open	
8000	5000	Closed	

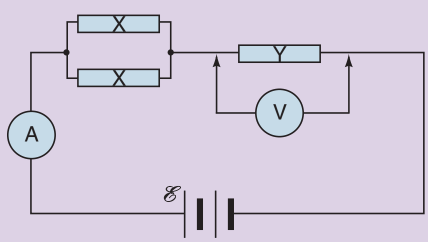
***12** For the circuit shown, find the:

a current in the 20 Ω resistor
b potential difference between points X and Y.



***13** Two electronic components X and Y, are operating in the circuit shown. The voltage–current graphs for each device are shown on the axes below.

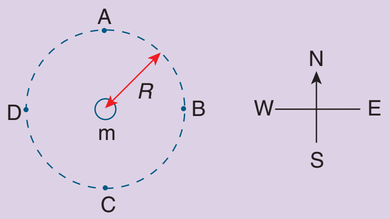




a Which device is non-ohmic?
b What is the resistance of X?
c What is the reading on the ammeter?
d Determine the EMF of the battery.
e Calculate the total power consumption in the circuit.

The following information applies to questions 14–20.

The diagram below shows a conductor m with its axis perpendicular to the plane of the page. The local magnetic flux density of the Earth is 5.00×10^{-5} T. A current, I, flowing through the conductor produces a magnetic field of magnitude 50.0 μT at a distance R.



14 In which direction (into or out of the page) would the current need to be flowing through this conductor for it to produce the following field directions at the points specified?
a east at A
b west at A
c south at D
d east at C

15 What is the magnitude and direction of the magnetic flux density produced by this conductor at point B, if the resultant field at this point is zero?

16 What is the direction of the electric current responsible for producing this field?

17 What is the magnitude and direction of the resultant magnetic flux density at point B for a current, I, flowing through this conductor out of the page?

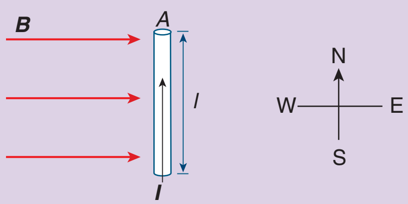
18 A current, I, flows through this conductor out of the page. What is the magnitude of the resultant magnetic flux density at point A?

19 What is the direction of the resultant magnetic field at point C for a current, I, flowing into the page?

20 This conductor is carrying a current, I, out of the page when the direction of the current is reversed. Which of the following correctly describes the *change* in the magnetic field at point B produced by this current reversal?
A 5.00×10^{-5} T south
B 1.00×10^{-4} T north
C 1.00×10^{-4} T south

The following information applies to questions 21–23.

The following diagram shows a conductor of length l carrying a current, I, perpendicular to a magnetic field of magnetic flux density $B_\perp$. Ignore the magnetic field of the Earth.



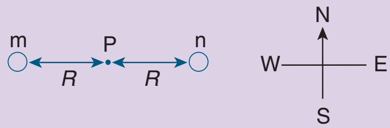
21 Calculate the magnitude and direction of the magnetic force on this conductor for the following sets of data:
a B = 1.00 mT, l = 5.00 mm, I = 1.00 mA
b B = 0.100 T, l = 1.00 cm, I = 2.00 A
c B = 1.00 T, l = 10.0 mm, I = 5.00 A

22 The direction of the magnetic field and current are changed. Describe the direction of the magnetic force on this conductor for:
a B west, I north
b B west, I south.

23 What would be the magnitude of the magnetic force on this conductor if it were aligned so that the current was parallel to the field?

The following information applies to questions 24–27.

The diagram shows two conductors, m and n, with their axes perpendicular to the page. A current, I, flowing through either conductor will produce a magnetic field of magnitude 1.00 T at point P.



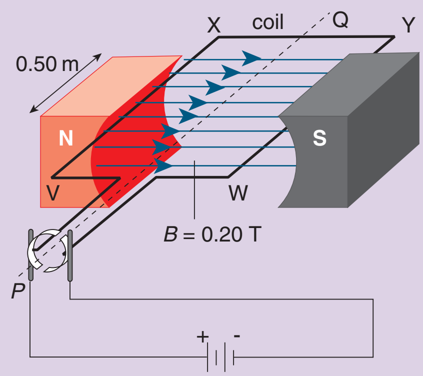
24 If a current, I, flows through both conductors, state the nature of the force (attraction or repulsion) between these conductors for:
a I into page for both conductors
b I out of page for both conductors
c I into page for m, I out of page for n

25 A third conductor, carrying a current, I, into the page, is placed at point P. What is the magnitude of the force per metre on this conductor if m and n both carry currents, I, into the page?

26 Conductor m now carries a current, I, into the page, while n carries a current, I, out of the page. What will be the magnitude of the force per metre on a third conductor placed at point P if this conductor carries a current I = 1.00 A?

The following information applies to questions 27–29.

The diagram shows a simplified version of a direct current motor.



27 For the position of the coil shown, calculate the magnitude of the force on segment WY when a current of 1.0 A flows through the coil.

28 In which direction will the coil begin to rotate?

29 Which of the following actions would cause the coil to rotate faster?
A increasing the current
B increasing the magnetic field strength
C increasing the cross-sectional area of the coil
D all of the above

4 Generating electricity

In this chapter, we explore electromagnetic induction—the creation of an electric current from a changing magnetic field. Whether the primary source of energy is burning coal, falling water, nuclear fission or the Sun, virtually all of the electricity generated in the world's power stations is the result of electromagnetic induction. In fact, electromagnetic induction is one of nature's fundamental principles. It has become the basis of our modern technological society, and it has enabled us to deepen our understanding of the fundamental processes of the Universe including the nature of light, the origin of the Earth's magnetic field, and much more.

In 1831, Englishman Michael Faraday and American Joseph Henry independently discovered how to create an electric current by using magnetism, beginning a massive expansion in our understanding and use of electricity. It became clear that it would be possible to produce electricity in quantities far beyond the capacities of chemical batteries. A new phase of the industrial revolution was about to begin.

By the end of this chapter

you will have covered material from the study of generating electricity, including:

- the generation of an EMF by electromagnetic induction
- the factors that give rise to an induced current in a coil in which there is a changing magnetic flux
- transformers—what they do, how they work
- alternating voltage and current
- electric power production and energy usage
- the transmission of electric power.

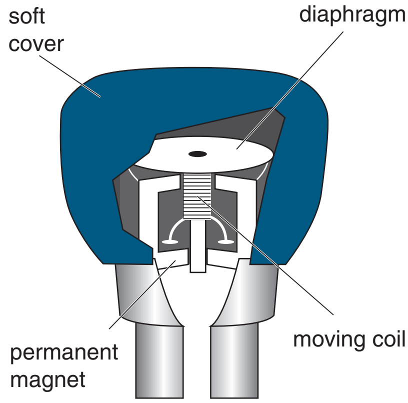


Figure 4.1
Sound waves vibrate the diaphragm of a microphone. These vibrations are converted to a tiny electric current by the relative motion of the coil and permanent magnet.

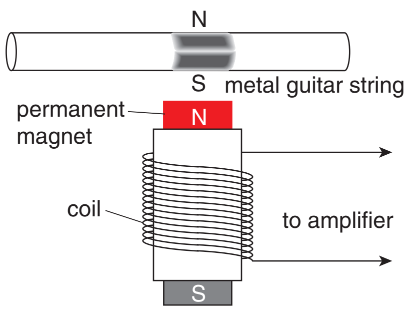


Figure 4.2
A guitar pick-up magnetises the steel string with a permanent magnet and then uses this vibrating magnet to induce an EMF in a coil below it.

4.1 Magnetic flux and induced currents

Microphones and electric guitars

The function of a microphone is to convert sound vibrations into electrical signals. The so-called 'dynamic' microphone uses a tiny coil attached to a diaphragm, which vibrates with the sound. The coil vibrates within the magnetic field of a permanent magnet, thus producing an induced EMF which varies with the original sound.

An acoustic guitar amplifies sound from the vibrating strings by setting up resonance in the air and the wood of the guitar body. The electric guitar, on the other hand, uses electromagnetic induction to generate an electric signal which is then amplified electronically. The pick-up consists of a coil around a small permanent magnet. The permanent magnet induces the section of the wire above it to become a magnet. As the string vibrates, this moving magnet induces an EMF in the coil. This is then amplified, and often electronically 'doctored', to produce the sound required.

Physics in action — Faraday's discovery

After Oersted's discovery that an electric current produces a magnetic field, Michael Faraday was convinced that somehow a magnetic field should be able to produce an electric current.

In one of his attempts to create an electric current with magnetism, Faraday wound two coils of wire onto an iron ring. Into one coil he fed a strong current from a battery, thus creating a strong magnetic field in the ring. But no matter how strong a magnetic field he managed to produce, he could not detect an electric current in the other coil. One day he noticed that his galvanometer gave a small kick when he turned the field on, and another kick, in the opposite direction, when he turned it off. It was not the presence of the field that mattered, but the fact that it changed!

He soon found that any method of *changing* the amount of magnetic field cutting through a coil of wire created a small current in the wire, but only while it was changing. Simply moving a permanent magnet into a coil induced a brief current pulse, with another brief pulse in the opposite direction when it was removed.

Figure 4.3
An early galvanometer of the type Faraday would have used in his studies of electricity and magnetic fields. A galvanometer is simply a sensitive current detector. Faraday used a fixed coil of wire around a magnetic needle. The needle responded to the field created by a current in the coil. Modern mechanical galvanometers use a coil of many turns free to rotate in a strong magnetic field.

The creation of an electric current in a loop of wire due to changes in a magnetic field is called ***ELECTROMAGNETIC INDUCTION***.

We can gain a feel for **electromagnetic induction** by experimenting with a coil of wire near a permanent magnet or an electromagnet. The coil is connected to a galvanometer—a sensitive current meter. If the magnet is moved towards the coil, the galvanometer registers a current, but only while the magnet is moving, not once it stops. If the magnet is moved away, a current in the opposite direction is registered. In fact it does not matter whether it is the coil or the magnet that is in motion—it is only the *relative* motion that is important.

Next, we use a stationary electromagnet but change the field by turning it on and off. When the current in the electromagnet is switched on, there is a brief pulse of current through the galvanometer, indicating an induced current in the coil as the magnetic field increases. While a steady current flows in the electromagnet, no current is registered by the galvanometer. When the electromagnet is turned off again, the galvanometer indicates another brief pulse of current in the coil, this time in the opposite direction. If we increase or decrease the current in the electromagnet (i.e. if we change the strength of the field) the galvanometer again registers a small current. So it seems that any method of changing the 'amount' of magnetic field cutting through the coil will induce a current. Indeed, even changing the shape of the coil so that more or less field passes through it results in an induced current. In all cases, the larger the change, and the faster the change, the greater the current.

To describe the 'amount of field' more precisely, physicists use a quantity called *magnetic flux* (Φ_B), defined as the product of the magnetic flux density, $\boldsymbol{B}$, and the area, $A_\perp$, of the loop perpendicular to the field lines.

MAGNETIC FLUX is given by:

$$\Phi_B = \boldsymbol{B}A_\perp$$

where Φ_B is the magnetic flux passing through the coil (Wb), $\boldsymbol{B}$ is the magnetic flux density (T) and $A_\perp$ is the area of the coil perpendicular to the direction of the magnetic field (m^2).

Faraday pictured a magnetic field represented by many *lines of force*, the closeness of the lines representing the strength of the field. Where the lines are crowded (near the poles of a bar magnet, for example) the field is strong; where they are less dense the field is weaker. Magnetic flux then, is rather like the total number of these lines—the same flux could be produced by a weak field spread over a large area (the lines are spread out) or a strong field concentrated in a small area (a high density of lines). This picture leads to the expression for the strength of the magnetic field, $\boldsymbol{B}$, as the *magnetic flux density*. $\boldsymbol{B}$ can be thought of as proportional to the number of lines of force passing through a unit of area that is perpendicular to the lines.

Thus, the unit of magnetic flux is the magnetic flux density (tesla) multiplied by area (m^2), or T m^2. This unit is also known as the weber (1 Wb = 1 T m^2). Conversely, this implies that the unit for magnetic flux density can be expressed as weber per square metre (Wb m^{-2}), a unit that is still often used.

Faraday went on to show that the amount of induced current in a coil was proportional to the rate at which the number of lines of force cutting

Figure 4.4
The original apparatus with which Faraday discovered electromagnetic induction.

Physics file

We use the term 'loop' for a single closed conducting path, such as a circle of wire, and the term 'coil' for a series of loops wound together.

Physics file

If the loop is not perpendicular to the field, then we need to find the component of the area of the coil that is perpendicular to the field and then multiply this by the magnetic flux density:

$$\Phi_B = \boldsymbol{B}A\cos\theta$$

where Φ_B is the magnetic flux passing through the coil (Wb), $\boldsymbol{B}$ is the magnetic flux density (T), A is the area of the coil (m^2) and θ is the angle between the field lines and the normal to the plane of the coil.

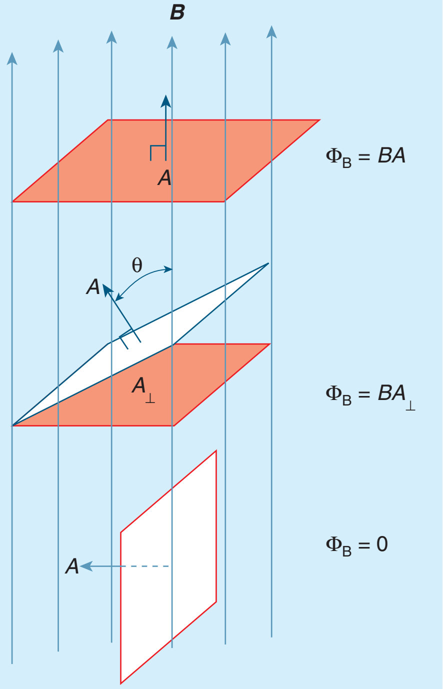


Figure 4.5
The magnetic flux is the strength of the magnetic field, $\boldsymbol{B}$, multiplied by the effective area ($A_\perp = A\cos\theta$), here shown shaded.

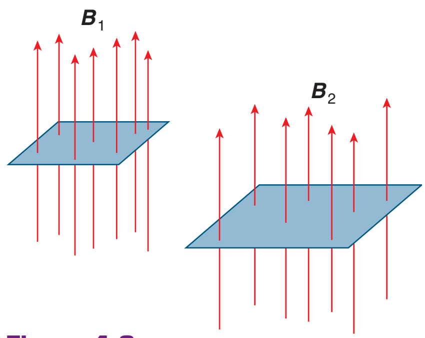


Figure 4.6
(a) A strong field. (b) A weaker field, but the same amount of flux

through a coil was changing; that is, the rate of change of magnetic flux passing through the coil.

> The induced current in a conducting loop is proportional to the rate of change of flux:
>
> $$I_{ind} \propto \frac{\Delta\Phi_B}{\Delta t}$$
>
> where I_{ind} is the conventional current induced in the loop (A), Φ_B is the magnetic flux passing through the component of the loop that is perpendicular to the field (Wb) and Δt is the time period over which the flux changes (s).

✓ Worked Example 4.1A

A student places a horizontal 5.00 cm square coil of wire into a uniform vertical magnetic field of 0.100 T.

a How much flux cuts through the coil?

She then pulls the coil out of the field at such a rate that it takes 1.00 s for the coil to lose all the flux. While the coil is coming out of the field, she finds that a current of +2.00 mA is registered on the ammeter connected to it.

b If she puts the coil back into the flux at the same rate, what current will she observe?

c If she pulls it out again, but twice as fast as before, what current will be induced?

Now she steadily rotates the loop while it is fully in the field in such a way that it takes 2.00 s to rotate through 180.0°.

d Describe the current that will be induced.

Solution

a $\boldsymbol{B} = 0.100$ T $\qquad \Phi_B = \boldsymbol{B}A_\perp = (0.100)(0.0500^2)$

$A_\perp = 0.0500^2\ \text{m}^2 \qquad = 2.50 \times 10^{-4}$ Wb

b The same current but in the opposite direction: −2.00 mA.

c The induced current is related to the change in flux over the change in time so:

$$I_{ind} \propto \frac{\Delta\Phi_B}{\Delta t}$$

This time Δt was halved so the current will be doubled. The current will be +4.00 mA.

d The total change of flux that occurs in the 2.00 s is:

$$DF_B = F_{B\,final} - F_{B\,initial}$$
$$= (-\ 2.50\cdot\ 10^{-4}) - (2.50\cdot\ 10^{-4})$$
$$= -\ 5.00\cdot\ 10^{-4}\ \text{Wb}$$

As the flux changes from $+2.50 \times 10^{-4}$ to -2.50×10^{-4} Wb, the sign of the flux is related to the side of the coil it passes through. As the time is doubled the average rate of this flux change is the same as in part a (-2.50×10^{-4} Wb s^{-1}), so the average current will again be ±2.00 mA. However, this will not be uniform. It will be a maximum (±2.00 mA) as the loop passes through the vertical and a minimum as the loop is horizontal. The magnitude of the current is not related to the magnitude of the flux, but to the magnitude of the rate of change of flux.

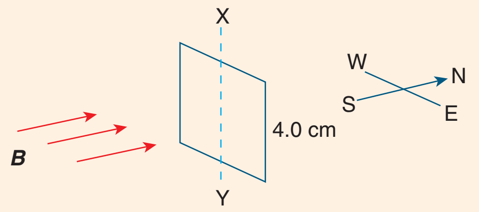
4.1 SUMMARY Magnetic flux and induced currents

- Symmetry suggests that if an electric current creates a magnetic field, it should be possible to use a magnetic field to create a current.
- It is relative movement between the field lines and the loop that induces a current in a conducting loop.
- Magnetic flux is defined as the product of the magnetic field and the perpendicular area over which it is spread; that is, $\Phi_B = \boldsymbol{B}A_\perp = \boldsymbol{B}A\cos\theta$ where θ is the angle between the field lines and the normal to the plane of the coil.
- Any way of changing the magnetic flux through a loop induces a current in the loop. The greater the rate of change of flux, the greater the current.

4.1 Questions

The following information applies to questions 1–3.

A student makes a square loop of wire, of side 4.00 cm, is in a region of uniform magnetic field, $\boldsymbol{B} = 2.00 \times 10^{-3}$ T north, as in the following diagram. The student then rotates the loop about a vertical axis XY. When the loop is in its initial position, its plane is perpendicular to the direction of the magnetic field and the angle θ between the plane of the loop and north is 0°.



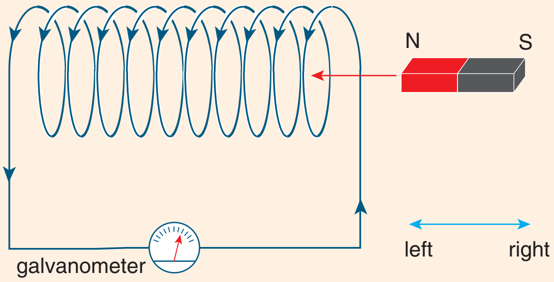
1 What is the magnetic flux passing through the loop when θ has the values 0.00°, 45.0°, 60.0° and 90.0°?

2 Calculate the magnitude of the change in magnetic flux through the loop when θ changes from 0.00° to 45.0°, 0.00° to 60.0°, 45.0° to 90.0°, and 0.00° to 90.0°.

3 The loop is fixed in its initial position (θ = 0.00°). Determine the change in magnetic flux through the loop when the following changes are made to the original magnetic field.
   a The magnetic field strength is reduced to zero.
   b The direction of the magnetic field is reversed.
   c The magnetic field strength is doubled.
   d The magnetic field strength is halved.

The following information applies to questions 4 and 5.

A coil of wire connected to a galvanometer forms a circuit, as shown in the following diagram. When a bar magnet is placed near the coil and moved to the left, the galvanometer indicates a positive current.

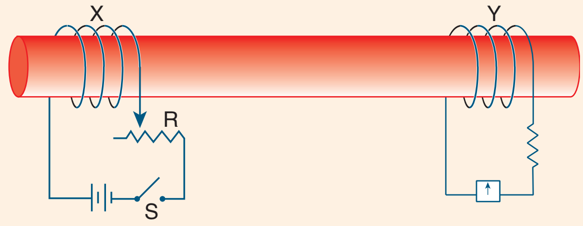


4 For each of the following situations state whether the current through the galvanometer will be zero, positive or negative.
   a coil stationary, magnet stationary
   b coil stationary, magnet moved to right
   c coil moved to right, magnet stationary
   d coil moved to left, magnet stationary

5 What condition is necessary for a magnetic field to induce a current in a coil?

The following information applies to questions 6 and 7.

Two coils X and Y are wound on the same length of steel rod, as shown. The magnitude of the current through coil X can be altered using a variable resistor R. The switch, S, is initially open. When S is closed, a positive current flows through coil X.

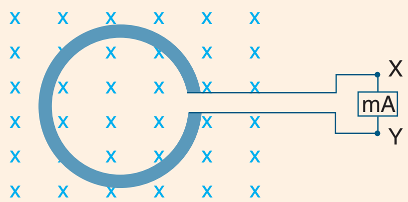


6 Describe the current flowing in coil Y as S is closed, held closed for some time, and then opened.

7 The switch is closed again and remains closed. Describe the current flowing in coil Y when the resistance is at first steadily increased and then steadily decreased.

The following information applies to questions 8–10.

A coil of radius 4.00 cm has its plane perpendicular to a uniform magnetic field of strength 2.00 mT directed into the page, as in the following diagram.



8 a What is the magnetic flux through the coil?
   b What is the current in the coil?

9 The magnetic field is then switched off in such a way that it takes 1.00 ms to drop to zero. When switched on again, it also takes 1.00 ms. Switching off the field results in a momentary current of 4.00 mA flowing through the milliammeter from X to Y.
   a What is the change in magnetic flux through the coil when the field is turned off?
   b What is the rate of change of flux through the coil during this time?
   c What momentary current will flow through the coil when the field is switched back on?

10 Find the induced current in the coil when the following changes are made to the same coil.
   a The direction of the original magnetic field is reversed in 1.00 ms.
   b The original coil is replaced with one of radius 2.00 cm and the original field is then switched off.
   c The original field is switched off but it takes 2.00 ms instead of 1.00 ms to decrease uniformly to zero.

4.2 Induced EMF: Faraday's law

Faraday realised that the induced current depended on the particular characteristics of the coil as well as the flux changes. He found that it was the EMF (or voltage) induced that was independent of the particular characteristics of the circuit. The reason for this is not hard to see if we remember that a current carrying wire that is perpendicular to a magnetic field will experience a force. This is just what is happening as a coil moves through a magnetic field.

First we will consider the simple case of a straight wire moving in a magnetic field. In the last chapter we saw that, when the positive charged conventional current flows through a conductor that is aligned parallel to an external magnetic field, then a force will move the wire. Once again symmetry suggests that if we move the charged particles contained within a conductor through a magnetic field current should result. This is the case. As a charge q moves through a perpendicular magnetic field $\boldsymbol{B}_{\perp}$ at a speed v, there is a magnetic force on it equal to $qv\boldsymbol{B}_{\perp}$. The right hand rule tells us that the force on positive charges in the wire in Figure 4.7 would be in the direction along the wire and out of the page. The force on the negative electrons in the wire will be inwards. Thus the closer end of the wire (to us) will become positive, and the other end, into the page, will become negative. This will result in a potential difference (ΔV) created across the ends of the wire.

Consider an electron moving through this wire, starting from the closer end and moving inwards. The $qv\boldsymbol{B}$ force will do work on the electron as it moves along the length of the wire in the field l, given by:

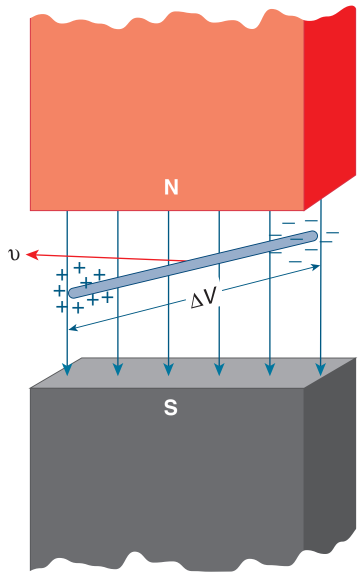


Figure 4.7
A potential difference ΔV will be produced across a straight wire moving to the left in a downward magnetic field.

Physics file

It does not matter that any one particular electron may not actually travel along the whole length of the wire. This same analysis applies to each little segment of wire, and the overall change of potential is the sum of all the changes in potential for each little segment.

$$W_d = \boldsymbol{F}l = qv\boldsymbol{B}_{\perp}l$$

where W_d is the work done on the charge (J), q is the charge of the particles within the conductor (C), v is the speed of the wire moving in the field (m s^{-1}), $\boldsymbol{B}_{\perp}$ is the perpendicular magnetic flux density through which the wire is moving (T) and l is the length of wire in the field (m).

Given that there are many other charges being moved as well, this work will go into the potential energy due to the concentration of positive charge at one end relative to the concentration of negative charge at the other. There is then a potential difference between the ends of the wire. If they were connected by a closed circuit outside the field, a current would flow.

The potential difference is equal to the potential energy gained per unit of charge. By dividing the previous expression by the charge q, we find:

$$\Delta V = \frac{W_d}{q} = \frac{qv\boldsymbol{B}_{\wedge}l}{q}$$

or

$$\Delta V = v\boldsymbol{B}_{\wedge}l$$

where ΔV is the potential difference or EMF, $\mathscr{E}$ (V), v is the speed of the wire moving in the field (m s^{-1}), $\boldsymbol{B}_{\perp}$ is the magnetic flux density through which the wire is moving (T) and l is the length of wire in the magnetic field (m).

✓ Worked Example 4.2A

As an aeroplane flies along through the Earth's magnetic field, it is acting rather like the straight wire described above.

a At which places on the Earth will the induced EMF across an aeroplane's wings be maximum?

b What is the maximum EMF which would be induced across the wings of a Boeing 747 with a wing span of 64.0 m, flying at 990 km h^{-1} through the Earth's magnetic field? Take the maximum magnitude of the Earth's magnetic field as 60.0 μT.

Solution

a As the plane flies horizontally, it cuts through the vertical component of the Earth's field. Near the magnetic poles, the field is close to vertical, so this is where we would expect the maximum induced EMF. Near the Equator the Earth's field is nearly horizontal and so the plane is flying parallel to the field lines and does not cut them; thus there will be no induced EMF along the wing.

b The maximum magnitude of the induced EMF near the poles is found from:

$v = \dfrac{990}{3.6}$ m s^{-1}

$B_\perp = 60.0 \times 10^{-6}$ T

$l = 64.0$ m

$$\text{EMF} = vB_\perp l = (275)(60.0 \times 10^{-6})(64.0) = 1.06 \text{ V}$$

Not only is this EMF very small, but any attempt to use it, or even measure it, would be thwarted by the fact that the same EMF will be induced in any wires connecting the wing tips, giving us two equal but opposite EMFs in the circuit.

The relationship $\mathscr{E} = vB_\perp l$ is only useful for a straight wire moving in a magnetic field. However, it can be used to derive a more generally useful expression if we note that, as the wire moves, it sweeps across lines of magnetic flux. In time Δt the wire moves a distance d, which is equal to $v\Delta t$, so the area ΔA it has swept out is:

$$\Delta A = l \times d = lv\Delta t$$

So the amount of magnetic flux that the wire sweeps through in this time interval is given by:

$$\Delta\Phi_B = B_\perp \Delta A = B_\perp lv\Delta t$$

This can be written:

$$B_\perp lv = \frac{\Delta\Phi_B}{\Delta t} = \frac{B_\perp \Delta A}{\Delta t}$$

But the EMF, $\mathscr{E}$, is equal to $vB_\perp l$, so:

$$\text{EMF} = \frac{\Delta\Phi_B}{\Delta t} = \frac{B_\perp \Delta A}{\Delta t}$$

So the magnitude of the EMF generated in the wire is equal to the rate at which the wire is sweeping through the magnetic flux.

While this derivation was for a straight wire moving across a field, it is also true whenever a change of flux gives rise to an EMF. A rigorous derivation (taking into account the vector nature of the quantities involved) leads to Faraday's law of induction.

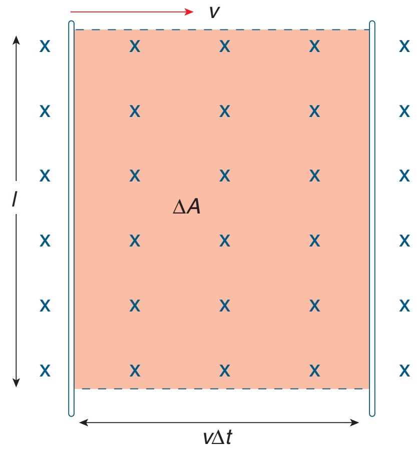


Figure 4.8
The area ΔA (shaded) swept out by the wire in the field in a time Δt is equal to $lv\Delta t$.

Physics file

The EMF, or electromotive force ($\mathscr{E}$), is a potential difference created by giving charges potential energy. In a battery the energy comes from chemical reactions. In this situation the energy comes from whatever is moving the wire (the steam turbine in a power station, for example). Despite the name, EMF is not a force, but a voltage.

FARADAY'S LAW OF INDUCTION states that:

$$\text{EMF} = -\frac{\Delta\Phi_B}{\Delta t}$$

That is, the **EMF GENERATED** in a conducting loop in which there is a changing magnetic flux is equal to the negative rate of change of flux.

Practical activity

31 Faraday's law of electromagnetic induction

Notice that a negative sign has appeared in the expression. This has an important significance (which will be explained later in this section), but it is normally ignored in simple calculations. This law is known as Faraday's law of induction in honour of its discoverer. This expression applies to a single conducting loop. Since the turns in a coil of wire are in series:

In a coil of N turns, the **TOTAL EMF** is given by:

$$\text{EMF} = -N\frac{\Delta\Phi_B}{\Delta t}$$

If the ends of the coil are connected to an external circuit, a current will flow. Ohm's law tells us that the current would be equal to $\frac{\Delta V}{R}$, where R is the total resistance in the circuit. This is consistent with the fact that an induced current always arises when the flux through a coil changes, and that this current is proportional to the rate of change of flux. A coil not connected to a closed circuit will act like a battery not connected in a circuit. There will be an EMF, but no current.

✓ Worked Example 4.2B

A student winds a coil of area 40.0 cm² with 20 turns. He places it horizontally in a vertical uniform magnetic field of 0.100 T, and connects it to a galvanometer with resistance 200.0 Ω.

a How much flux passes through the coil?

b If it is then withdrawn from the field in a time of 0.500 s, what would be the average current reading on the galvanometer?

Solution

a $B_\perp = 0.100$ T

$A = \frac{40.0}{100 \times 100}$ m²

$\Phi_B = B_\perp A$

$= (0.100)(4.00 \times 10^{-3})$

$= 4.00 \times 10^{-4}$ Wb

b The average EMF induced in each turn is given by the average rate of change of flux:

$\Delta\Phi_B = 4.00 \times 10^{-4}$ Wb

$\Delta t = 0.500$ s

$N = 20$ turns

$\text{EMF} = -N\frac{\Delta\Phi_B}{\Delta t}$

$= -(20)\frac{4.00 \times 10^{-4}}{0.500}$

$= -1.60 \times 10^{-2}$ V

Assuming that the only significant resistance in the circuit is the galvanometer, the average induced current is given by Ohm's law:

$\Delta V = 1.60 \times 10^{-2}$ V

$R = 200.0\ \Omega$

$I = \frac{\Delta V}{R}$

$= \frac{1.60 \times 10^{-2}}{200.0}$

$= 8.00 \times 10^{-5}$ A

The direction of EMF: Lenz's law

It is important to know the direction in which the induced current will flow in a circuit. In fact, the negative sign in Faraday's law implies a direction, but the vector analysis required is difficult. Fortunately there is a simpler way of finding the direction. Just 3 years after Faraday's discovery, Heinrich Lenz (a German physicist working in Russia) discovered a simple principle by which the direction of the induced EMF could be found. He realised that the principle of conservation of energy meant that any magnetic flux produced by the induced current must tend to oppose the original *change of flux*, which gave rise to the induced current.

LENZ'S LAW states that any induced current in a loop will be in the direction so that the flux it creates will oppose the ***change*** in the flux that produced it.

It is not hard to see that this has to be the case. Consider the loop shown in Figure 4.10a. The magnet is moving to the left and increasing the amount of flux in the loop. Clearly a current will be induced in the loop, but in which direction? Imagine that the induced current is clockwise (upwards in the near side of the loop). Use of the right-hand grip rule (as in Figure 4.9) will show that this current will produce a magnetic flux that points in the same direction as that from the magnet. Now this would create an interesting situation! The flux from the induced current would add to the increasing flux from the magnet and produce a greater change of flux. This in turn would increase the induced current, which would increase the flux change, which would further increase the induced current—and so on. In other words, we have an impossible situation. A small initial change of flux would lead to an ever-increasing induced current. Clearly this cannot be allowed, as we would be obtaining electrical energy from nowhere!

Lenz realised that the flux created by the induced current must be such as to *oppose* the *change of flux* from the magnet. If we imagine the reverse of the previous situation—that is, an induced current that flows anticlockwise around the loop (Figure 4.10b)—we can see that the flux created by the induced current does oppose the increasing flux from the magnet. This is indeed the situation we find in practice—the induced current creates a flux that reduces the actual change of flux in the loop.

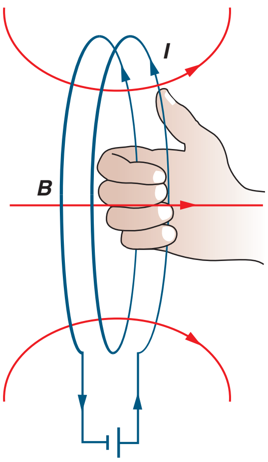


Figure 4.9
Using the right-hand grip rule to find the direction of flux from the current in a coil, or vice versa

Physics file

How does the Earth create its magnetic field? The energy that drives the Earth's dynamo comes from the enormous heat produced by radioactive decay deep in the Earth's core. The heat causes huge swirling convection currents of molten iron in the outer core. These convection currents of molten iron act rather like a spinning disk. They are moving in the Earth's magnetic field and so eddy currents are induced in them. It is these eddy currents that produce the magnetic field.

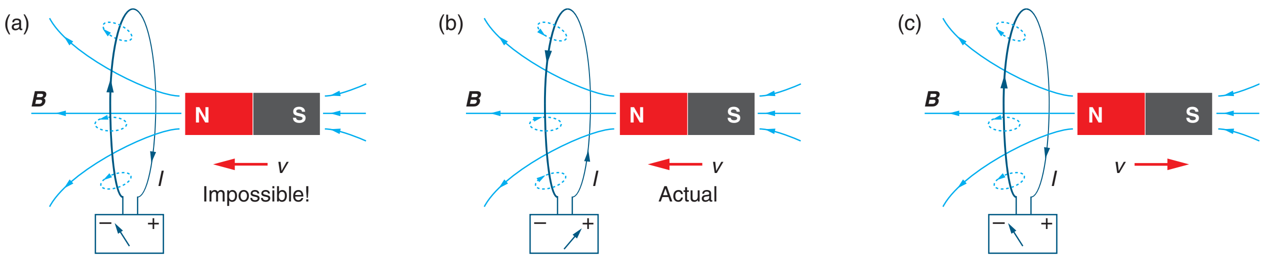


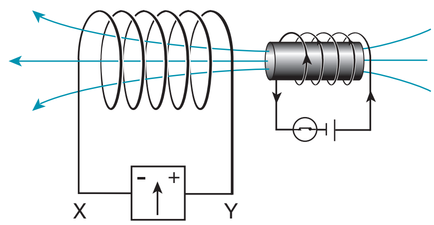
Figure 4.10
(a) An impossible situation! As the magnet approaches, the induced current would rapidly escalate, contravening the principle of conservation of energy. (b) The actual induced current opposes the changing flux. (c) As the magnet is withdrawn, the induced current creates a flux, which opposes the loss of flux.

If the magnet is moved away from the loop instead of towards it, the flux, while still pointing to the left, is decreasing (Figure 4.10c). In this situation, as you might expect, the induced current creates a flux that points in the same direction as the flux from the magnet. Again, it opposes the change of external flux by creating a flux that attempts to replace the decreasing flux.

✓ Worked Example 4.2C

Instead of using a permanent magnet to change the flux in the loop in Figure 4.10, an electromagnet could be used. What is the direction of the current induced in the loop when the electromagnet is:

a switched on?
b left on?
c switched off?



Solution

a When the electromagnet is switched on, the induced current must create a flux which will oppose the increasing flux to the left, just as in the case of the permanent magnet earlier. Thus the current will flow through the meter in the direction from Y to X.
b While the current in the electromagnet is steady, the flux is not changing and therefore there will be no induced EMF or current in the loop.
c When the electromagnet is switched off, the flux is decreasing and hence the induced current will now flow through the meter in the direction from X to Y.

To create an induced current we do not always need loops and magnets. Whenever a changing magnetic flux encounters a conducting material an induced current will occur. These currents are often called *eddy currents* and may result in lost energy in electrical machinery. On the other hand, the so-called 'Faraday dynamo' utilises this effect to create large currents. A copper disk is rotated in a strong perpendicular magnetic field. Application of the right-hand palm rule will show that a current will flow from the centre to the rim of the disk (or vice versa). Because of the very low resistance involved, very large currents can be generated and drawn off from contacts at the centre and rim of the disk.

The eddy currents produced in a moving conductor will themselves be subject to the ***IlB*** force. This has the effect of retarding the motion that is giving rise to the current and so can be used as a magnetic brake. Many car speedometers utilise eddy currents. A magnet, connected by a cable to the wheels, spins close to an aluminium disk to which the speedo pointer is attached. Eddy currents induced in the disk experience a force that tends to drag the disk around against a spring—the faster the magnet, the greater the force. A more dramatic use of eddy currents is in maglev trains, which float on a repulsive magnetic force between magnets in the train and eddy currents in the special track.

You may wonder how there can ever be an increase in the flux in a conducting loop if the induced current always sets up a flux that opposes any change. If the loop were a perfect conductor, it would not be possible, but in fact most loops and coils are not perfect conductors and, as a result of their resistance, the induced currents die out quickly. However, at very low temperatures, some materials are perfect conductors and have zero resistance. Such conductors are called **superconductors**. If a magnet is brought near a superconductor, the induced current will create a flux which, within the superconductor, will exactly cancel the flux from the magnet. A similar effect enables a very strong magnet to be made from a superconducting ring.

Physics file

Eddy currents in action: find a small strong magnet, preferably a neodymium 'super' magnet, and drop it down a tube made of a good conductor such as copper or aluminium. You will observe the effects of eddy currents. See if you can deduce the direction of the eddy currents produced and hence reason out an explanation for the slow fall of the magnet.

Figure 4.11
A maglev train is supported by magnetic forces created by induced, or eddy, currents. Various systems are being trialled in several countries. Generally the force is between currents induced in flat conductors under the train by strong, possibly superconducting, magnets mounted in the fast-moving train—normal suspension is needed for slow travel. There are two components to the force: one is an undesirable drag force we normally notice when a conductor is moved slowly near a magnet; the other is the repulsive lift force. Fortunately, the drag force decreases with speed and the lift force increases with speed. At around 300 km h^{-1} the lift force is about 50 times stronger than the drag force.

Physics in action — Superconductors and superconducting magnets

Technological breakthroughs have often led to advances in physics. This was the case in 1908 when Kamerlingh Onnes, at the University of Leiden in the Netherlands, succeeded in liquefying helium. Helium liquefies at 4.2 K (−268.9°C). It was known that the electrical resistance of metals decreases as they cool, so one of the first things that Onnes and his assistant did was to measure the resistance of some metals at these very low temperatures.

Onnes was hoping to find that as the temperature of mercury dropped towards absolute zero its resistance would also gradually drop towards zero. What they found, however, was a complete surprise. At 4.2 K its resistance vanished completely!

Onnes coined the word 'superconductivity' to describe this phenomenon. Soon he found that some other metals also became superconducting at extremely low temperatures: lead at 7.2 K and tin at 3.7 K, for example. Curiously, metals such as copper and gold, which are very good conductors at normal temperatures, do not become superconducting at all. Onnes was awarded the 1913 Nobel Prize in Physics for his work in low-temperature physics.

Much of the great promise of superconductivity has to do with the magnetic properties of superconductors. In a superconductor an induced current does not fade away! As the resultant field opposes the changing flux, the magnet is repelled. This gives rise to the 'magnetic levitation' effect that is by now well known (Figure 4.12). On a large scale this could perhaps one day provide us with virtually frictionless maglev (magnetic levitation) trains.

Unfortunately, the superconducting metals lost their superconductivity in magnetic fields around 0.1 T—quite a moderate field. However, in the 1940s it was found that some alloys of elements such as niobium had higher 'critical temperatures' and, more particularly, retained their properties in stronger magnetic fields. By 1973 the niobium–germanium alloy Nb_3Ge held the record with a critical temperature of 23.2 K in a critical field of 38 T—an extremely strong field.

In 1986 an entirely new and exciting class of superconductors was discovered. Georg Bednorz and Karl Müller, working in Switzerland, found that compounds of some rare earth elements and copper oxide had considerably higher critical temperatures. They received the 1987 Nobel Prize in Physics for their work.

These new 'warm superconductors' are ceramic materials made by powdering and baking the metal compounds. Most ceramics are insulators; it was a combination of good science and inspired guesswork that led Müller to try such unlikely candidates for superconductivity. So far, the record is held by bismuth and thallium oxides with a critical temperature around 125 K—still rather cold, but significantly above the temperature of readily available liquid nitrogen (77 K).

Superconductivity, particularly in the newer materials, is still not fully understood. It can really only be discussed in terms of quantum physics, but one rather picturesque way of thinking about it is that electrons pair up and 'surf' electrical waves set up by each other in the crystal lattice of the material.

The promise of superconductivity is enormous: low-friction transport, no-loss transmission of electricity, and smaller and more powerful electric motors and generators. Superconducting magnets might be used to contain the extremely hot plasma needed to bring about hydrogen fusion, producing almost pollution-free energy in much the same way that the Sun does. There are, however, many difficulties to be overcome before these promises can be realised.

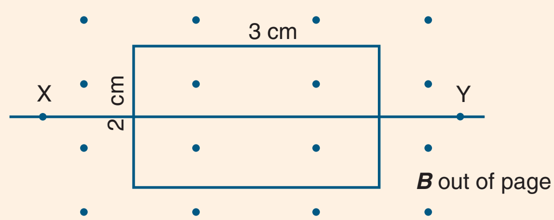
Figure 4.12
A disk magnet is repelled by a superconductor because the magnet induces a permanent current into the superconductor, which results in an opposing field.

4.2 SUMMARY Induced EMF: Faraday's law

- Faraday found that the EMF induced in a loop was dependent only on the rate of change of magnetic flux through the loop.
- As a conductor moves perpendicular to a magnetic field, work is done on the moving charges to produce a potential difference along the wire.
- This potential difference appears as an induced EMF in any loop in which the flux is changing.
- The induced EMF in a loop is equal to the (negative) rate of change of flux through the loop.
- Lenz's law states that the direction of the induced current in a loop will be such that the flux it creates will oppose the change in the flux that produced it.
- Lenz's law is an expression of the principle of conservation of energy.
- Induced currents set up by the relative motion of a conductor and a magnet will create a field that will apply a force that will oppose the relative motion.

4.2 Questions

1 A rectangular wire loop is located with its plane perpendicular to a uniform magnetic field of 2.00 mT, directed out of the page, as shown. The loop is free to rotate about a horizontal axis XY and has a resistance of 1.50 Ω.



a How much magnetic flux is threading the loop in this position?

b The loop is rotated about XY, through an angle of 90.0°, so that its plane becomes parallel to the magnetic field. How much flux is threading the loop in this new position?

c Calculate the average EMF induced in the loop if this rotation took 40 ms.

d What is the average current induced in the loop?

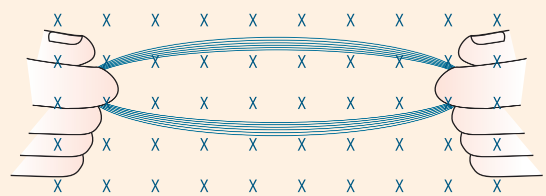
2 A coil of 500 turns, each of area 10.0 cm², is wound around a square frame. The plane of the coil is initially parallel to a uniform magnetic field of 80.0 mT. The coil is then rotated through an angle of 90° so that its plane becomes perpendicular to the field. The rotation is completed in 20.0 ms.

a What is the change in magnetic flux through each turn of the coil during this time?

b What is the average EMF induced in each turn during the rotation?

c Calculate the average EMF induced in the coil during this time.

3 A student stretches a flexible wire coil of 30 turns and places it in a uniform magnetic field of strength 5.00 mT, directed into the page, as shown. While it is in the field, the student allows the coil to regain its original shape. In doing so, the area of the coil changes at a constant rate from 50.0 cm² to 250.0 cm² in 0.50 s.



a What is the change in magnetic flux through each turn of the coil during this time?

b Find the average EMF induced in the coil during this time.

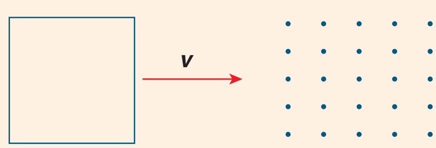
4 The vertical component of the Earth's magnetic field in Albany is 4.42×10^{-5} T upwards. A helicopter at Albany Airport has metal rotor blades 4.00 m long from axis to tip.

a How much magnetic flux would a rotor blade cut through during one rotation?

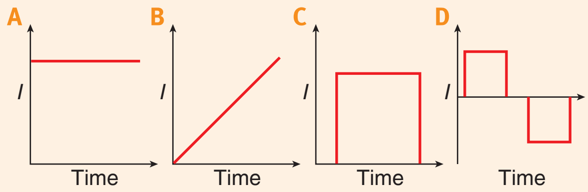
b If the blade rotates at 8.00 revolutions per second, calculate the average EMF induced between the axis and tip.

The following information applies to questions 5 and 6.

A square loop of conducting wire is moved with constant velocity ***v*** from a region of zero magnetic field into a region of uniform magnetic field and then out again into a field-free region.



5 Which of the graphs A–D best represents the current I in the loop as a function of time t?

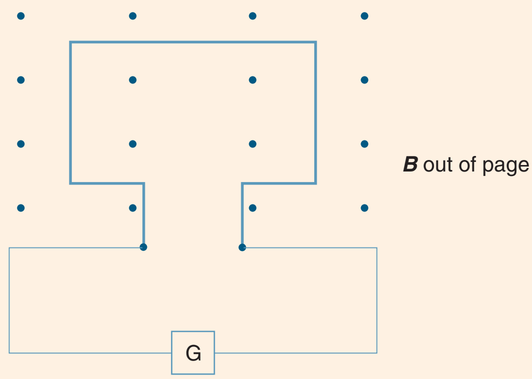


6 If the area of the loop is 1.6×10^{-3} m² and $v = 2.5$ m s⁻¹, calculate the magnitude of the magnetic field that could induce an EMF of 5.0 mV in the loop during its motion.

7 A conducting loop is located in an external magnetic field whose direction (but not necessarily magnitude) remains constant. A current is induced in the loop. Which of the following alternatives best describes the direction of the magnetic field due to the induced current?

A It will always be in the same direction as the external field.
B It will always be in the opposite direction to the external field.
C It will be in the direction that will oppose the change that produced the induced current.

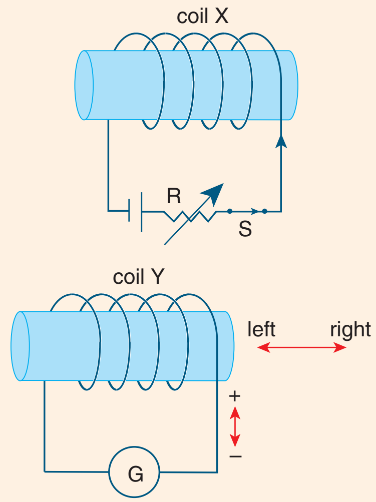
8 A rectangular conducting loop forms the circuit shown below. The plane of the loop is perpendicular to an external magnetic field whose magnitude and direction can be varied. The initial direction of the field is out of the page.



What is the direction of the magnetic field due to the induced current when the following changes are made to the initial field?

a The field is switched off.
b The field strength is doubled.
c The direction of the field is reversed.

9 Two coils X and Y are located next to each other. Coil X is connected to a circuit containing a battery, variable resistor R and switch S. Coil Y is connected to a circuit containing a galvanometer. Initially S is closed, producing a positive current through X and a magnetic field which extends through coil Y.



State the direction of the induced current through the galvanometer, when the following changes are made to the initial circuit containing X.

a S is opened.
b R is increased.
c R is reduced.

4.3 Electric power generation

> **Physics file**
>
> Whenever a current is induced in a wire moving in a magnetic field, there will be a force on it which will oppose the motion. This can be seen in Figure 4.8 (section 4.2): as the wire moves to the right, any current induced will flow towards the top of the page. This current will in turn experience a force to the left, i.e. in the direction opposite to the motion. We find this to be a general principle. If the movement of a conductor results in an induced current, the magnetic force on the current will oppose the motion. Because of this, energy must always be used to move a conductor in a magnetic field. It is this energy that will appear as electrical energy in the circuit.

We take the supply of electric power to our homes for granted. However, it is only a little over 100 years since the streets of Perth were first lit by electric light. At that time there were many small electric power companies, all supplying different voltages and different systems, some by alternating current and some by direct current.

The main battle between AC and DC systems took place in the United States of America, where Thomas Edison's General Electric Company championed a DC system while George Westinghouse promoted an AC system. In 1888 Westinghouse enlisted the brilliant young Serbian inventor Nikola Tesla, who had recently demonstrated his AC induction motor. This was to be the turning point in the battle. With the new motor, as well as other Tesla inventions, AC systems began to dominate. Nowadays, with the exception of train and tram systems (which convert the AC supplied by the power companies into DC), all major power supply systems are based on alternating current.

The basic principle of electric power generation is the same, whether the result is alternating or direct current: relative motion between a coil and a magnetic field induces an EMF, and hence a current, in the coil. In some generators, a coil is rotated in a magnetic field, but in large power stations, and in car alternators, the coils are stationary and an electromagnet rotates inside them.

It is important to remember that the energy appearing in the external circuit has to be provided by some mechanical source, such as steam, water or wind. This is, of course, a consequence of the tremendously powerful principle of nature we call the *law of conservation of energy*. If it were somehow possible to induce a current without expending energy, this principle would have been violated.

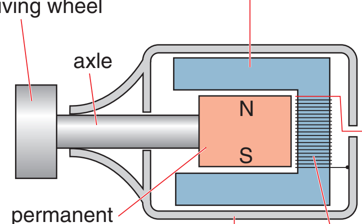


Figure 4.13

A bicycle dynamo is a simple rotating magnet generator. Because the current is taken from fixed coils there are few moving parts and little friction. The 'magneto' which powers the spark plug in small two-stroke engines uses a similar principle.

Electric power generators

Although most generators work on the rotating magnet principle, it is easier to consider the EMF induced in a coil rotating in a uniform magnetic field. In fact, the EMF produced in a rotating magnet generator has the same characteristics; it is only the relative motion that is important.

A rotating coil generator looks very like a standard DC motor. Energy is used to rotate a coil in a magnetic field, provided by either a permanent magnet or an electromagnet. The direction of the EMF induced in the coil (and therefore the current in the coil) will alternate as the flux through the coil increases and decreases. Consider the simple case of a rotating loop in a constant, uniform magnetic field, as shown in Figure 4.14.

The amount of flux cutting through the loop varies as it rotates. Remember that it is the *changing* field that induces the EMF. While the flux decreases from the maximum (a) to zero (b) and then becomes negative (c), Lenz's law tells us that the induced current will be in such a direction as to create a field in the same direction, relative to the loop, as the initial field. The right-hand grip rule then shows us that the current will flow in the direction D $\rightarrow$ C $\rightarrow$ B $\rightarrow$ A.

The induced current will change direction every time the flux reaches a maximum, i.e. when the plane of the loop is perpendicular to the field.

> **Physics file**
>
> Lenz's law tells us that the induced current creates a field to oppose the *change* of flux. This allows Faraday's law to conform to the law of conservation of energy. While the flux is decreasing, the field resulting from the induced current will therefore be in the same direction as the original field in order to try to maintain the flux in that direction.

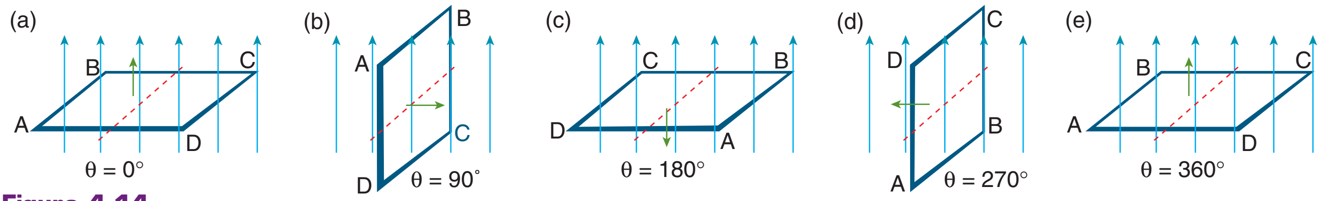


Figure 4.14

(a) The field $\boldsymbol{B}$ and the plane of the area A of the loop are perpendicular and the amount of flux through the loop is maximum $\Phi_B = \boldsymbol{B}A$. (b) After the loop has turned through a quarter of a revolution the plane of the loop is parallel to the field and the flux through the loop is zero. (c) The flux then increases but in the opposite sense relative to the loop; $\Phi_B = -\boldsymbol{B}A$. It then decreases to zero again (d). This is followed by another maximum (e) and the cycle repeats. The green arrow indicates the normal to the plane of the area. The angle θ is between the field and the normal to the plane.

To find the magnitude of the induced EMF, it is necessary to calculate the rate at which the flux through the loop is changing. The actual flux though the loop is given by:

$$\Phi_B = \boldsymbol{B}_{\perp}A$$

where Φ_B is the magnetic flux (Wb), $\boldsymbol{B}_{\perp}$ is the magnetic flux density (T) and A is the area of the loop normal to the magnetic field (m^2).

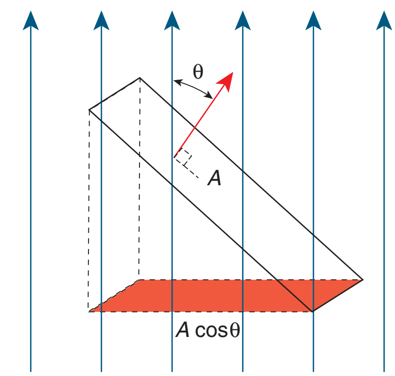


Figure 4.15

The angle θ is the angle between the field and the normal to the plane of the coil. It is usual to describe the orientation of a plane by the direction of the normal, as otherwise two directions would be required.

Interactive tutorial

AC generators

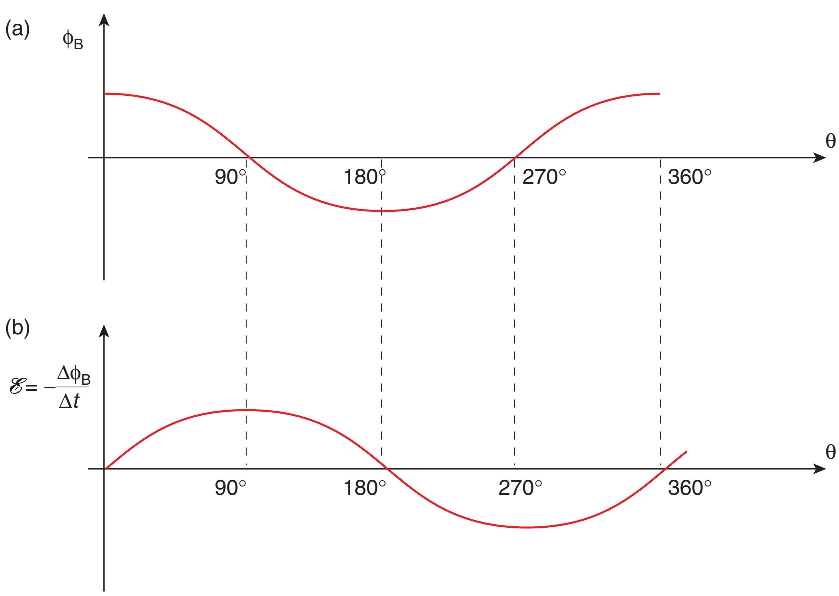


Figure 4.16

(a) The flux Φ_B through the loop as a function of the angle θ between the field and the normal to the loop. (b) The rate of change of flux $\Delta\Phi_B/\Delta t$ through the loop (and hence the EMF, $\mathscr{E}$) as a function of the angle θ between the field and the normal to the loop. The loop is rotating at a steady speed.

As we can see from Figure 4.16, the maximum rate of flux change, and hence EMF, occurs where the loop itself is parallel to the field and the flux through the loop is zero (i.e. where θ is 90° or 270°).

The actual value of the EMF generated will depend on the rate at which the flux is changing, as Faraday's law tells us. To find the expression for the EMF we simply divide the change in flux by the period of time over which it is changing. That expression is for a single turn loop. For a coil of N turns the EMF will be N times as much.

Physics file

An alternative notation for Faraday's law can be derived from a basic understanding of calculus. The flux through a loop turning in a uniform magnetic field is given by $\Phi_B = BA\cos\theta$. The EMF induced is the rate of change of this flux. The value of θ, the angle of the loop, can be expressed as ωt where ω is the 'angular velocity', the angle (in radians) turned through in one second. For a generator turning at 50 Hz the angular velocity will be:
$50 \times 2\pi = 100\pi$ radians s^{-1}
(2π radians is one full rotation or 360°)
Hence, the expression for the flux becomes $\Phi_B = BA\cos\omega t$. The rate of change of this flux is given by:

$$\mathscr{E} = \frac{d\Phi_B}{dt} = -\frac{d(BA\cos\omega t)}{dt} = BA\omega\sin\omega t$$

Practical activities

34 Hydroelectric power
35 Wind power

■ ***INDUCED EMF***, or ***FARADAY'S LAW***:

$$\text{EMF} = \frac{-N\Delta\Phi}{\Delta t}$$

where EMF is the electromotive force (V), N is the number of turns in the coil, $\Delta\Phi$ is the change in magnetic flux (Wb) and Δt is the period of time over which the flux changes (s).

Typically the change in flux is calculated from a quarter turn of the loop. The change in flux is calculated by finding the final flux minus the initial flux. This is best found by assigning maximum flux as the initial, and zero flux as the final. When this change occurs over a quarter turn of the loop the period of time (Δt) will be a quarter of the time it takes to spin the loop around once.

The current induced in the coil is drawn off by a commutator, which can consist of either two simple slip rings with brushes or a split cylinder commutator as in the DC motor described in Chapter 3. In the first case, the output will be just the alternating current which results from the EMF shown in Figure 4.17a, and the device is often referred to as an 'alternator'. In the second, the direction of the output is changed by the commutator every half turn and so the output current is always in the same direction (i.e. DC) although it varies from zero to maximum twice every cycle. This time, the device is called a DC generator (Figure 4.17b). The DC generator was widely used in cars up until the 1980s when semiconductor diodes enabled the simple conversion of AC into DC.

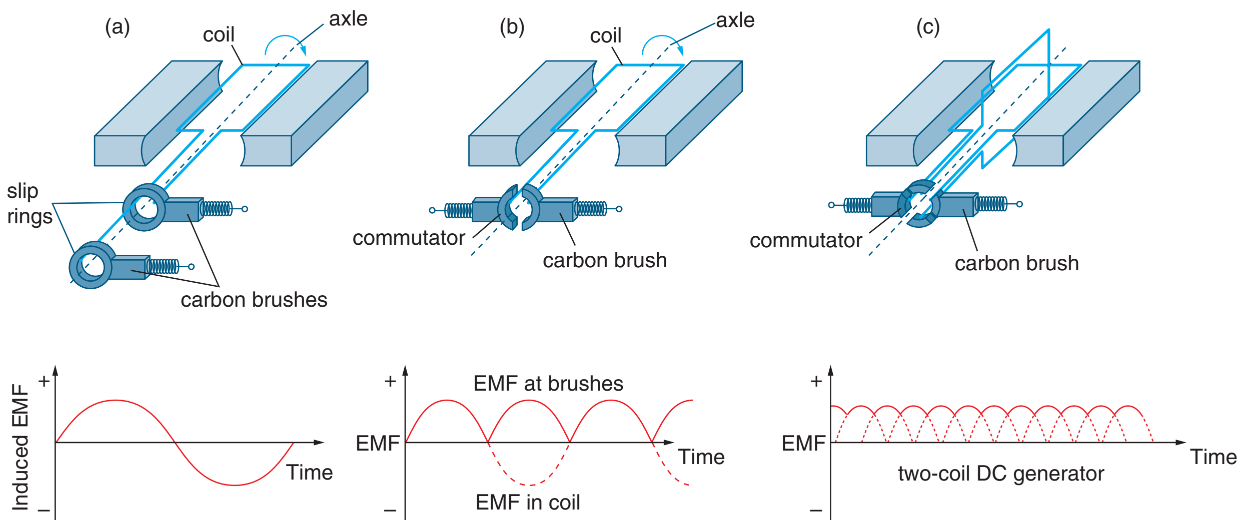


Figure 4.17
(a) An AC generator uses simple slip rings to take the current from the coil. (b) A DC generator has a commutator to reverse the direction of the alternating current every half cycle and so produce a DC output. (c) Two (or more) coils may be used to smooth the DC output.

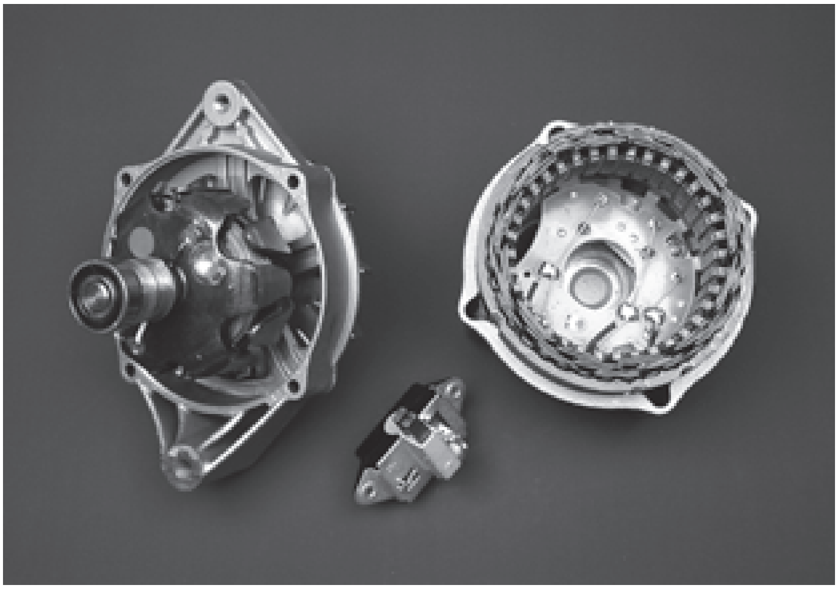


Figure 4.18

A modern car alternator employs a rotating electromagnet which induces an alternating EMF in the stator coils. This AC is often rectified by diodes inside the housing and so the output from the device is actually DC.

Alternating voltage and current

Nearly all electric power generators produce alternating current; that is, a current whose direction oscillates back and forth in a regular way at a certain frequency. Mains power in Australia oscillates at 50 cycles per second (Hz) and reaches a peak voltage of about ±340 V each cycle. Radio waves are produced from alternating currents of frequencies that range from less than 1 million hertz (1 MHz) to over 10 billion hertz (10 GHz). In this section, we look at some of the basic characteristics of AC electricity.

We can describe an AC voltage by the simple expression $V = V_p \sin\theta$ where V_p is the peak voltage (340 V for mains power) and θ varies from 0 to 360° every cycle. (θ is the angle of the rotor in the generator.) The *peak-to-peak* voltage is also sometimes quoted (e.g. $V_{p\text{-}p}$ = 680 V for mains power). Generators in Australian electricity supply systems rotate at 50 revolutions each second (3000 rpm), so the frequency of the alternating voltage produced is 50 Hz.

The current that flows through a simple resistive device, such as a light bulb, can be calculated by using Ohm's law. Thus, the current during the cycle will be given by $I = I_p \sin\theta$ where $I_p = V_p / R$.

Power in AC circuits

The power in any circuit element is the product of the current and the voltage, $P = IV$. In an AC circuit then, $P = I_p V_p \sin^2\theta$, or $P = (V_p^2 / R)\sin^2\theta$. As we might expect, the power varies at a frequency twice that of the alternating voltage (as $\sin^2\theta$ is +1 when $\sin\theta = \pm 1$). Physically, this can be seen in terms of the alternating movement of the electrons in the circuit. Whether they are moving one way or the other, they still transfer energy to the atoms of the conductor. We do not normally notice this 100 Hz variation of power in our lights as our eyes do not respond to flickering at frequencies greater than about 20 Hz.

Physics file

The main disadvantage of the DC generator is the commutator. Because it involves breaking the current there is the possibility of sparking. The sparking causes wear on the brushes and can gradually burn out the commutator. The alternator does not require the reversal of current and therefore can use slip rings which, while still creating some friction, do not involve breaking the current at all. In a rotating magnet alternator the slip rings only need to carry the relatively small DC current to the electromagnet. For these reasons modern car alternators are more reliable than their older DC counterparts.

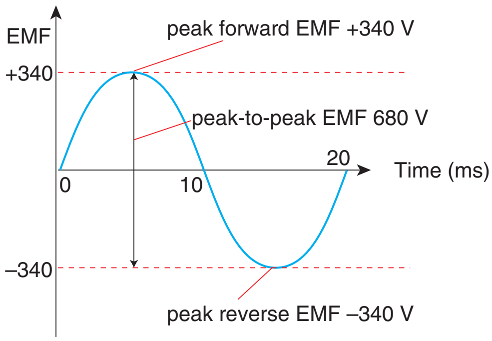


Figure 4.19

The voltage in our power points oscillates between +340 V and –340 V 50 times each second. The value of a DC supply that would supply the same average power is 240 V.

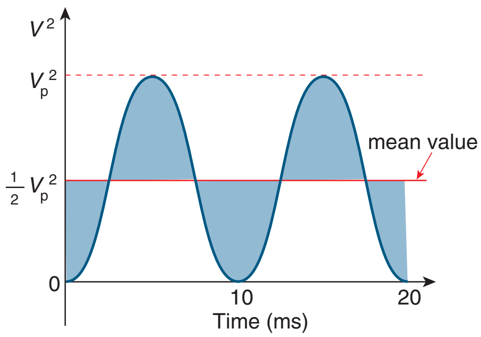


Figure 4.20

The square of the voltage for an AC supply. The average value of V^2 is equal to ½V_p^2. Note that the frequency is 100 Hz.

Physics file

In an AC circuit, the power produced in a resistor is equal to:

$$P = \left(\frac{V_p^2}{R}\right)\sin^2\theta$$

As can be seen from Figure 4.20, the *average power* will be given by:

$$P_{av} = \frac{V_p^2}{2R}$$

If this same power was to be supplied by a steady (DC) source, the voltage (say V_{av}) of this source would have to be so that:

$$\frac{V_{av}^2}{R} = \frac{V_p^2}{2R}$$

which simplifies to:

$$V_{av}^2 = \frac{V_p^2}{2}$$

$$V_{av} = \frac{V_p}{\sqrt{2}}$$

Because of the process of obtaining it, this voltage is known as the *root mean square* voltage or V_{RMS}. It is the value of a steady voltage which would produce the same power as an alternating voltage with a peak value equal to $\sqrt{2}$ times as much.

It is often more useful to know the *average* power produced in the circuit. As is shown in the adjacent Physics file, the average power can be obtained by using a value for the voltage which is equal to the peak voltage divided by √2. This is referred to as the *root mean square* or RMS voltage.

$$V_{RMS} = \frac{V_p}{\sqrt{2}}$$

$$V_p = \sqrt{2}V_{RMS}$$

where V_{RMS} is the average or root mean squared voltage (V) and V_p is the peak voltage (V).

In effect, the RMS voltage is the value of a DC voltage that would be needed to provide the same average power as the alternating voltage. In a simple resistive circuit, the current is directly related to the voltage ($I = VR$), and so a similar relationship will hold for the current:

$$I_{RMS} = \frac{I_p}{\sqrt{2}}$$

$$I_p = \sqrt{2}I_{RMS}$$

where I_{RMS} is the average or root mean squared current (A) and I_p is the peak current (A).

The mains voltage supplied to our houses has a peak value of 340 V. It is the RMS value of the voltage, $\frac{340}{\sqrt{2}} = 240$ V, that is normally quoted. This is the 'effective average value' of the voltage—the value which can be used to find the actual power supplied each cycle by an AC power supply.

Figure 4.21

A Tesla coil produces very high frequency alternating voltages. Although the voltage is very high, the frequency is so fast that the charges do not have time to enter the body before they are on the way out again; hence, it is relatively safe.

✓ Worked Example 4.3A

A 60.0 W light bulb supplied with 240 V AC uses 60.0 J every second, but the instantaneous power varies from 0.00 to 120.0 W, 100 times every second. Justify this statement.

Solution

The minimum power is obviously zero at the two points in every cycle when the voltage, and therefore the current, is zero. The quoted 60.0 W is the RMS power. It is obtained as the product of the RMS voltage and RMS current. The RMS voltage is 240.0 V and so the RMS current is given by:

$P = 60.0\ \text{W}$

$V_{RMS} = 240.0\ \text{V}$

$$I_{RMS} = \frac{P}{V_{RMS}} = \frac{60.0}{240.0} = 0.250\ \text{A}$$

The peak power is given by:

$I_{RMS} = 0.250\ \text{A}$

$V_{RMS} = 240.0\ \text{V}$

$$P_p = \sqrt{2}V_{RMS} \times \sqrt{2}I_{RMS} = 2(240.0)(0.250) = 120\ \text{W}$$

Physics in action — Three-phase electricity

All of the high-voltage transmission lines we see coming from power stations, or substations, have three conductors. The 240 V power lines in the street have at least three conductors. Only the lines running into individual houses have two.

Power generated in large power stations is *three-phase* power. In designing a large generator, it would be inefficient use of space to simply have one pair of diametrically opposite coils around the rotating magnet. More particularly, it would result in uneven forces on the rotor as it turned through one cycle. For these reasons three pairs of coils at 120° to each other are used, as shown in Figure 4.22. One end from each coil is connected to a common point termed the 'neutral'. The other ends of the coils are the three output phases which are one-third of a cycle apart.

It is these three phases that are carried by the three active conductors in the high-voltage power lines we see coming from the power stations and terminal stations. The common connection of the coils, or neutral, is *grounded* (literally, connected to the Earth) at the power station. In fact, very little current flows in the neutral because the currents in the three phases always tend to balance each other as they are flowing in opposite directions (see Figure 4.22b). At any time the average current is actually zero. Typically, three-phase electricity is used to run heavy electrical machinery in industrial applications, and sometimes domestic applications too, such as air conditioners and pool pumps.

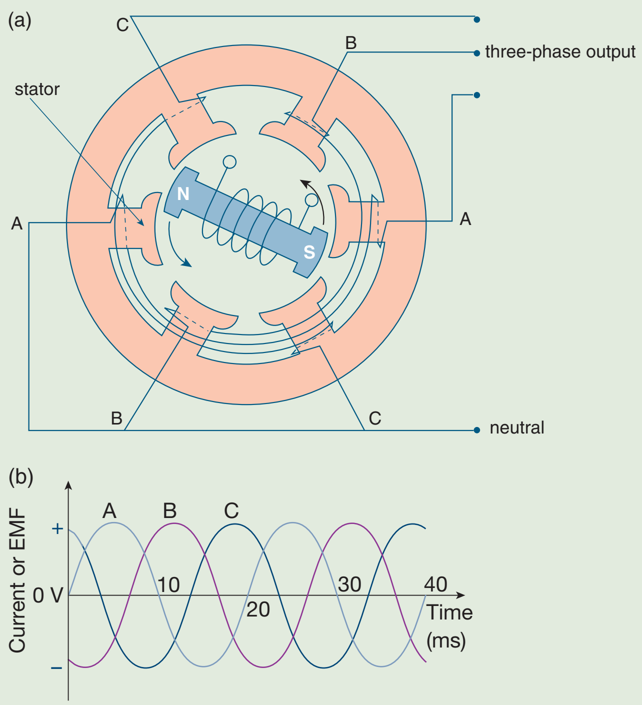


Figure 4.22
(a) A three-phase generator. The DC current is fed to the rotating magnet via slip rings (not shown). The ends of each of the three stator coils are connected together—this becomes the 'neutral'. The other ends carry the high AC voltages which are one-third of a cycle apart, as shown in (b).

4.3 SUMMARY Electric power generation

- An induced current in a loop will experience a force that will oppose the motion causing the current.
- The work done by the force moving the loop is equal to the electrical energy produced.
- The electrical output of a coil rotating in a magnetic field is sinusoidal. The peak EMF is proportional to the strength of the field, the area of the coil, the number of turns and the speed of rotation.
- A practical generator rotates either a coil inside a magnetic field or, more commonly, a magnet (permanent or electromagnetic) inside a fixed set of coils.
- The alternating current (AC) produced by power stations and supplied to cities varies sinusoidally at a frequency of 50 Hz. The peak value of the voltage of domestic power (V_p) is ±340 V, and $V_{p\text{-}p} = 680$ V.
- The root mean square voltage (V_{RMS}) is the value of an equivalent steady voltage (DC) supply which would provide the same power:

$$V_{RMS} = \frac{V_p}{\sqrt{2}}$$

The RMS value of domestic mains voltage is 240 V.

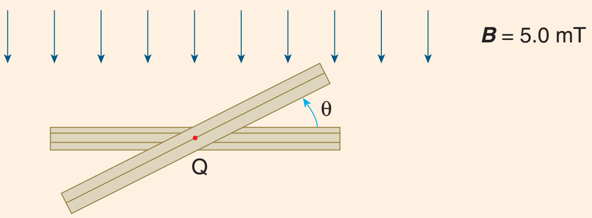
- For simple resistive circuits the current is given by Ohm's law, and $I_{RMS} = \frac{I_p}{\sqrt{2}}$
- The average power in a resistive AC circuit is given by:

$$P = V_{RMS} \times I_{RMS} = \tfrac{1}{2} V_p I_p$$

4.3 Questions

The following information applies to questions 1–6.

The following diagram shows a square coil containing 100 turns each of area 20.0 cm², mounted on a horizontal axis through Q. The plane of the coil is initially perpendicular to a vertical uniform magnetic field of 5.00 mT. The coil is to be rotated about Q in a counter-clockwise direction at a rate of 15.0° per millisecond (i.e. a frequency of 42 Hz).



1 What is the magnetic flux (Φ_B), in mWb, at each of the following values of θ?
   a 0.00° b 15.0° c 30.0° d 45.0°
   e 60.0° f 75.0° g 90.0°

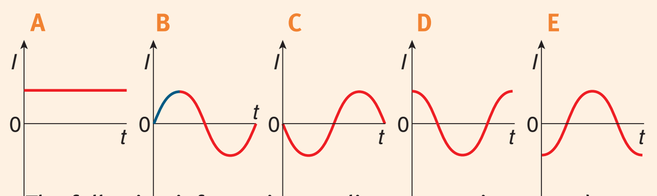
2 Calculate the average rate of change of flux through the coil when it is rotated through each of the six 15.0° intervals from 0.00° to 90.0° as in Question 1.

3 Considering your answers to Question 2, what do you notice about the rate of change of flux through the coil as it rotates from θ = 0.00° to θ = 90.0°?

4 Determine the average EMF induced in the coil during each of the first six 1.00 ms intervals of its rotation.

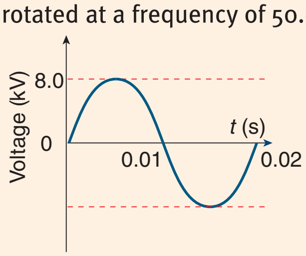
5 a At what value of θ do you think the peak value of induced EMF would occur? Justify your answer.
   b What is the peak value of the induced EMF for this coil?

6 Assuming that an anticlockwise rotation of the coil from θ = 0.00° initially produces a positive current, which of the graphs best illustrates the variation of the induced current as a function of time?



The following information applies to questions 7 and 8.

A simple generator consists of a coil of $N = 1000$ turns each of radius 10.0 cm, mounted on an axis in a uniform magnetic field of strength B. The following graph shows the voltage output as a function of time when the coil is rotated at a frequency of 50.0 Hz.



7 Which of the graphs below best describes the voltage–time relationship of this generator for the following modifications?

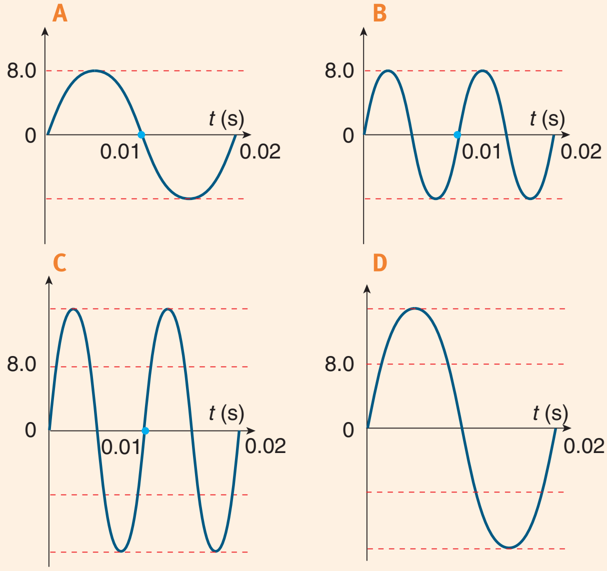
a $N = 1000$, $r = 10.0$ cm, field $= B$, $f = 100.0$ Hz

b $N = 1000$, $r = 10.0$ cm, field $= 2B$, $f = 50.0$ Hz

c $N = 500$, $r = 10.0$ cm, field $= 2B$, $f = 100.0$ Hz

d $N = 1000$, $r = 5.00$ cm, field $= 2B$, $f = 100.0$ Hz

e $N = 2000$, $r = 10.0$ cm, field $= B$, $f = 50.0$ Hz



8 Calculate the strength of the magnetic field required to produce a peak voltage of 8.00 kV.

9 An electrician is called to check a household power point and confirms that the voltage between the active (A) and neutral (N) is 240 V RMS.

a What is the peak voltage between A and N?

b What is the peak-to-peak voltage between A and N?

c An appliance of total resistance 100 Ω is plugged into the socket. What is the peak current that will flow in the circuit?

d What is the RMS current that will flow in the circuit?

10 An electric toaster designed to operate at 240 V RMS has a power rating of 600 W.

a What is the resistance of the heating element in the toaster?

b What is the peak voltage across the heating element?

c What is the peak current in the heating element?

4.4 Transformers

When Faraday first discovered electromagnetic induction, he had virtually invented the **transformer**. Any transformer is simply two coils wound so that the magnetic flux from one goes through the other. A common iron core increases this 'flux linkage' considerably.

Although a detailed analysis of transformer operation is complex, the basic idea is simple enough. Figure 4.24 shows an 'ideal' transformer. The two coils are wound on one core so that all the magnetic flux generated by one passes through the other. The coil connected to the AC supply is referred to as the *primary*, and the coil connected to the 'load' is the *secondary*.

A transformer operates on the principle that, whenever a changing magnetic flux passes through a coil, there will be an induced EMF. In a transformer, the changing flux originates from an alternating current in the primary coil. Because this changing flux also goes through the secondary coil, an EMF will be induced in that coil.

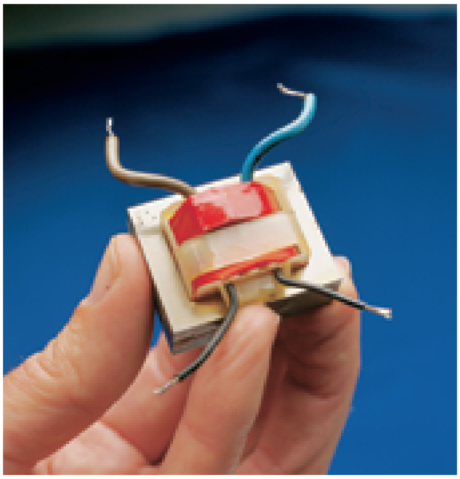


Figure 4.23
A modern transformer is most commonly made by winding the two coils around an iron core.

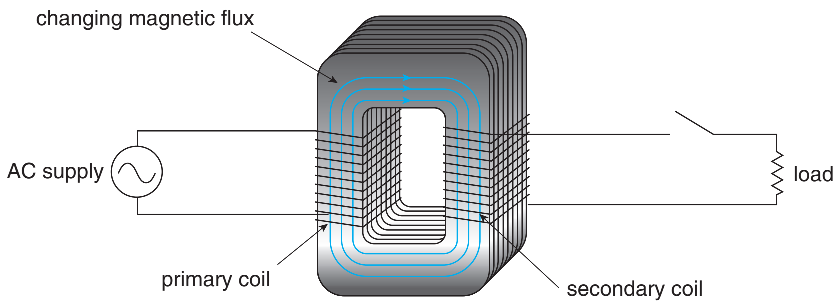


Figure 4.24

In an ideal transformer, the iron core ensures that all the flux generated in the primary also passes through the secondary.

Consider first the case where there is no load connected to the secondary coil. No current flows in the secondary and so all the flux is being generated by the primary coil. As we have seen, however, this flux will induce a voltage in the primary coil that will oppose the change of flux. That is, as the current increases, this 'back EMF' will tend to reduce the current. In a good transformer, this process is very effective and very little current will flow in the primary if none is flowing in the secondary.

Practical activity

33 Transformer operation

As very little current flows in the primary, we can see that the back EMF in the primary must almost be equal to the applied (mains) voltage. Also, as the same voltage must be induced in each turn of both coils (it is equal to the rate of change of flux), we can deduce that the voltage induced in the secondary coil will depend on the number of turns it has. Therefore, if it has the same number of turns as the secondary, it will have the same voltage. In any case, the ratio of the voltage induced in the secondary to that of the primary will be equal to the ratio of turns in each coil:

$$\frac{V_p}{V_s} = \frac{N_p}{N_s}$$

where V_p is the potential difference supplied to the terminals of the primary coil (V), V_s is the voltage or potential difference induced across the terminals of the secondary coil (V), N_p is the number of turns in the primary coil and N_s is the number of turns in the secondary coil.

Even when the load draws a current from the transformer, the voltage ratio does not change much. Provided we do not overload the transformer, it will maintain a secondary voltage very close to that given by the ratio of the turns. A transformer with a greater number of primary turns than secondary turns is referred to as a *step-down* transformer and one with a greater number of secondary turns is a *step-up* transformer.

A transformer effectively transfers EMF from one coil to another and does not use energy itself—apart from small resistive power losses in the windings and eddy currents in the core. Assuming these losses to be negligible, the law of conservation of energy enables us to assume that all the energy that goes into the primary will be transferred to the secondary. (A well-designed transformer might lose around 1–2% of the electrical energy that passes through it.) The rate of energy transfer is the power, so we can say:

Physics file

Because very little current can flow into the primary of a good transformer to which there is no load connected, it is possible to leave a transformer connected to the supply permanently and it will use very little power. In many of the electronic devices we leave on 'standby', transformers are constantly connected to the mains but, because of the back EMF, very little current is used while the device is not operating. However, over the whole community this can add up to megawatts of wasted power!

$$V_p I_p = V_s I_s$$

$$\frac{I_s}{I_p} = \frac{V_p}{V_s}$$

where V_p is the potential difference supplied to the terminals of the primary coil (V), V_s is the voltage or potential difference induced across the terminals of the secondary coil (V), I_p is the current in the primary coil (A) and I_s is the current in the secondary coil (A).

The ***TRANSFORMER EQUATIONS*** are:

$$\frac{I_s}{I_p} = \frac{V_p}{V_s} = \frac{N_p}{N_s}$$

✓ Worked Example 4.4A

A transformer used for charging a mobile phone battery operates on 12.0 V from the mains 240.0 V supply.

a If the number of turns in the secondary coil is 100, what will be the number of turns in the primary coil?

b If the battery charging requires a current of up to 4.00 A, what will be the current and the power drawn from the mains?

Solution

a

$N_s = 100$ $\qquad \frac{N_p}{N_s} = \frac{V_p}{V_s}$

$V_s = 12.0$ V $\qquad N_p = \frac{N_s V_p}{V_s} = \frac{(100)(240)}{12.0}$

$V_p = 240.0$ V $\qquad = 2000$

b The current ratio is the inverse of the turns and voltage ratio so that the current in the primary will be:

$N_s = 100$ $\qquad \frac{N_p}{N_s} = \frac{I_s}{I_p}$

$I_s = 4.00$ A $\qquad I_p = \frac{I_s N_s}{N_p} = \frac{(4.00)(100)}{2000}$

$V_p = 240.0$ V $\qquad = 0.200$ A

$V_s = 12.0$ V $\qquad P = VI$

$I_s = 4.00$ A $\qquad = (12.0)(4.00)$

$\qquad = 48.0$ W

The power required by the battery was 48.0 W. Alternatively, the input power was 48.0 W, which is of course the same, as we assumed that the transformer itself uses no power.

Physics file

A transformer will be overloaded if too much current is drawn and the resistive power loss in the wires becomes too great. Because this loss increases with the square of the current ($P = I^2R$), there will be a point at which the transformer starts to overheat rapidly. For this reason, it is important not to exceed the rated capacity of a transformer.

Physics file

Eddy currents set up in the iron core of transformers can generate a considerable amount of heat. Energy has been lost from the electrical circuit and the transformer may become a possible fire hazard. To reduce eddy current losses, the core is made of laminations—thin plates of iron electrically insulated from each other and placed so that the laminations interrupt the eddy currents.

Physics in action — AC induction motors

In order to have a current flowing in the armature of a DC motor, it is necessary to use brushes which rub on the commutator, with consequent wear and friction. Tesla realised that a current could be induced in a rotor without any mechanical contact by using, in effect, the transformer principle. His idea was simple: he used the changing magnetic flux from stationary coils, fed with AC current, to induce a current into a solid conducting rotor (see Figure 4.25a). That, however, was not sufficient to cause the rotor to move.

Tesla's trick was to make the field appear to rotate. The alternating current in the stator coils made the field appear to be continually reversing direction, just as though it was rotating. The current induced in the rotor of the induction motor opposed the relative motion between the apparently rotating field and the rotor—and so the rotor tried to spin with the field!

Of course, the field could seem to be rotating in either direction, but Tesla distorted the field by placing a conducting ring on one side of the pole, as in Figure 4.25b. This asymmetry made the field appear to rotate in one direction.

A great advantage of three-phase power is that it is just what is needed to make the magnetic field always appear to rotate in one direction. In the three-phase induction motor, three pairs of coils, each connected to one of the three phases, are arranged at 120° intervals. As the alternating current in each coil peaks, so the field appears to travel from one to the next, creating in effect a rotating field.

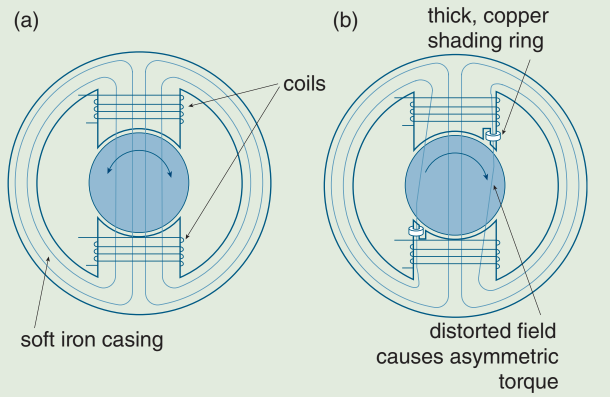


Figure 4.25
The principle of a simple single-phase, 'shaded pole' induction motor. The distorted (or 'shaded') field causes the rotor to turn in one direction in preference to the other.

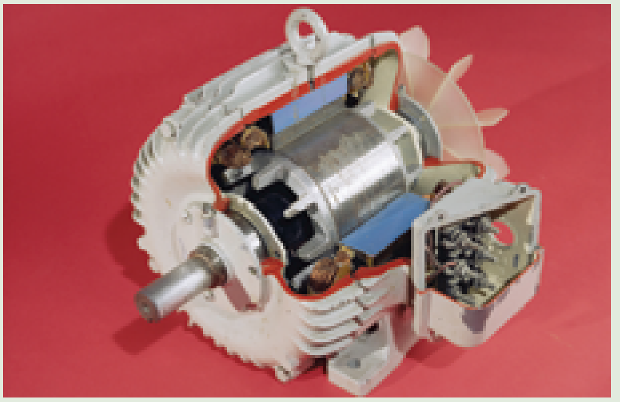
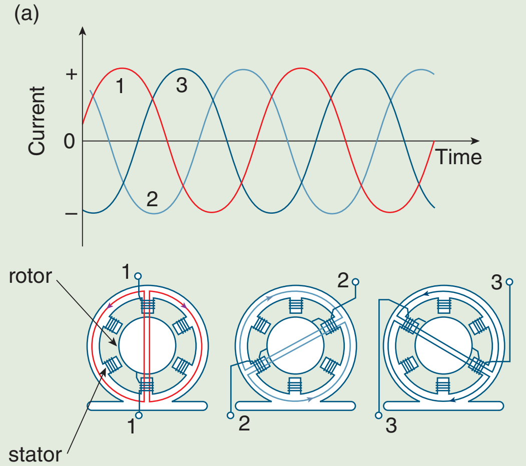


Figure 4.26
A typical AC induction motor. Induction motors are the most common electric motor used in industry. Because they have no brushes they are both more efficient and more reliable.



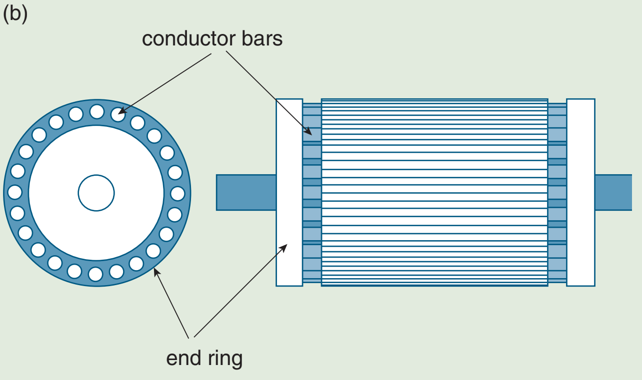


Figure 4.27
(a) In a three-phase motor, as the current in each pair of opposite coils peaks, the field appears to rotate, dragging the rotor around with it. (b) A 'squirrel cage' rotor. The rotor is made of iron laminations to cut down undesirable eddy currents. The induced currents flow lengthwise in copper or aluminium rods which are joined at the ends (as in a squirrel cage).

4.4 SUMMARY Transformers

- A transformer consists basically of two coils wound on an iron core so that all the magnetic flux generated by one also passes through the other.
- Where there is no load on the secondary a 'back EMF' in the primary opposes the current and reduces it to almost zero.
- Each turn in both the primary and secondary coils experiences the same flux changes and so the voltage ratio is given by:

$$\frac{V_p}{V_s} = \frac{N_p}{N_s}$$

- Assuming no power loss in the transformer, the power into the primary is the same as the power out of the secondary. Thus the current ratio is the inverse of the turns and voltage ratio:

$$P_p = P_s$$

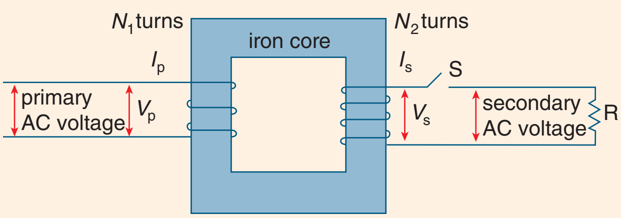
$$V_p I_p = V_s I_s$$

$$\frac{I_s}{I_p} = \frac{V_p}{V_s}$$

4.4 Questions

The following information applies to questions 1–6.

The diagram depicts an iron core transformer. An alternating voltage applied to the primary coil produces a changing magnetic flux $\Delta\Phi_B/\Delta t$. The secondary circuit contains a switch S in series with a resistor, R. The number of turns in the primary coil is N_1 and in the secondary N_2.



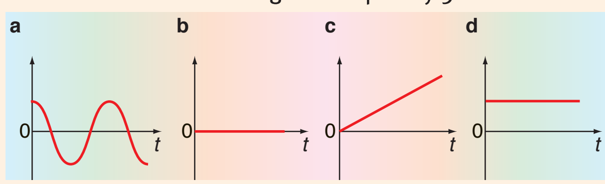
1 Assuming that the transformer is ideal, write an equation that defines the relationship between:
   a V_p and $\Delta\Phi_B/\Delta t$
   b V_s and $\Delta\Phi_B/\Delta t$
   c V_p and V_s

2 With S closed, which one or more of A–D equals the:
   a input power to the primary coil of an ideal transformer?
   b output power from the secondary coil of an ideal transformer?
      A $V_p I_p$
      B $V_s I_s$
      C $V_p I_s$
      D $I_s^2 R$

3 Again with S closed, which one or more of A–D correctly describes the:
   a power input to the primary coil of a non-ideal transformer?
   b power output from the secondary coil of a non-ideal transformer?
      A $V_p I_p$
      B $V_s I_s$
      C $V_p I_s$
      D $I_s^2 R$

4 What are the sources of power loss in a non-ideal transformer?

5 A cathode ray oscilloscope is now connected to the output of the transformer, and a series of different inputs are used. Which of the graphs (A–D) is the most likely output displayed on the CRO for:
   a a steady DC voltage?
   b a DC voltage that is increased gradually from zero at a constant rate?
   c a sinusoidal voltage of frequency 50 Hz?



6 Assume that the primary winding consists of 20 turns and the secondary of 200 turns. The primary RMS voltage input is 8.00 V and a primary RMS current of 2.00 A is flowing.
   a What is the RMS voltage across the load in the secondary circuit?
   b What RMS current will flow in the secondary circuit?
   c What power is being supplied to the load?

The following information applies to questions 7–9.

A security light is operated from mains voltage (240 V RMS) through a step-down transformer with 800 turns on the primary winding. The security light operates normally on a voltage of 12.0 V RMS and a RMS current of 2.00 A. Assume that the light is operating normally and that there are no losses in the transformer.

7 a Calculate the number of turns in the secondary coil.
   b What is the value of the peak current in the primary coil?
   c Calculate the power input to the primary coil of the transformer.

8 During some routine maintenance work the primary coil is unplugged from the AC mains and mistakenly connected to a DC supply of 240.0 V. Would the security light still operate? Justify your answer.

9 The primary coil is reconnected to the mains supply. It is then decided that the globe in the security light needs replacing and it is removed. This results in no current flowing in the secondary circuit for 10 minutes, during which time the primary coil is still connected to the supply. How much energy is consumed by the primary circuit during this time? Justify your answer.

4.5 Distributing electricity

Electric power for our cities

Modern cities use huge amounts of electrical power, which mostly has to be supplied from a long way away. Whether from a coal-fired power plant, which must be near a source of coal, a wind farm on the coast, or a nuclear power station, which must be near a large river or the sea for its cooling water, the transmission lines are often many hundreds of kilometres long.

POWER LOSS in transmission lines:

$$P_{loss} = \Delta VI \quad \text{and} \quad \Delta V = IR$$

$$P_{loss} = IR \times I$$

$$= I^2R$$

also

$$P_{loss} = \Delta V \times \frac{\Delta V}{R}$$

$$= \frac{\Delta V^2}{R}$$

where P_{loss} is the rate at which electrical energy is converted to other less useful forms of energy (W), I is the current, when using AC current it represents the RMS current (A), ΔV is the potential difference, with AC electricity it represents the RMS potential difference (V), and R is the resistance of the transmission line (Ω).

The transmission of electrical energy over large distances is therefore a very important consideration for power engineers, particularly in a country like Australia, with its widely separated population centres. A large city such as Perth uses up to 3000 MW of power at peak times. To provide a potential of 240 V at the end of the transmission lines this would require a current of 1.25×10^7 A (12.5 million amperes). No practical conductor could carry this current over long distances, so how can the power be transmitted from the generating station to the users?

The solution is to use very much higher voltages. The power arriving at the end of a transmission line is the product of the current carried and the potential at the end of the line ($P = VI$). The higher the potential, the lower the current needed. At a potential of 500 kV, the 3000 MW could be carried by a current of 6.0×10^3 A, or 6 000 A. This is much more feasible. Any practical power line has a significant resistance, which causes a power loss (see Physics file, this page). Clearly it is important to reduce the current as much as possible, as the power loss depends on the square of the current. For example, if the current were doubled, the resistance would have to be reduced by a factor of four to avoid more power loss, so conductors four times the weight would be required.

Physics file

The resistance of a material depends on the length and cross-section as well as the nature of the material. This is summed up in the expression:

$$R = \frac{\rho L}{A}$$

where R is the resistance in ohms (Ω), L is the length (m), A is the cross-section area (m^2) and ρ is the *resistivity* in ohm metres (Ω m).
Resistivity can be thought of as a measure of the inherent resistance of the material. The resistivity of copper is 1.72×10^{-8} Ω m and that of aluminium is 2.65×10^{-8} Ω m.

Physics file

There is a fundamental reason that 500 kV is about the limit for power transmission. The intense electric field near a small point at a high voltage can ionise the air in the vicinity, allowing it to conduct the charge away. This *corona effect* can sometimes cause a faint glow in the dark. The lightning rod, used on buildings prone to lightning strikes, utilises this principle—the corona discharge can neutralise the clouds. The field around a single very high voltage cable would also result in a corona discharge. At voltages of up to around 500 kV this can be avoided by using conductors consisting of four cables held apart by spacers. The four spaced conductors have the effect of creating a larger conductor which reduces the intensity of the electric field (just as the dome of the Van de Graaff generator does). The corona effect makes transmission voltages greater than this impracticable.

✓ Worked Example 4.5A

At a time of peak demand 835 MW is to be provided to Perth from the Muja Power Station, along a power line with a total resistance of 2.00 Ω.

a What would be the total transmission power loss if the potential at Perth was to be:

i 240 V? **ii** 500 kV?

b What potential would be needed at the Muja Power Station end of the line to achieve these potentials at Perth?

c How would the answers to parts a (ii) and b change if the resistance of the power line was halved?

Solution

a i

$P_{\text{Perth}} = 635 \times 10^6 \text{ W}$

$V_{\text{Perth}} = 240 \text{ V}$

$$I = \frac{P_{\text{Perth}}}{V_{\text{Perth}}} = \frac{635 \times 10^6}{240}$$
$$= 2.65 \times 10^6 \text{ A}$$

Thus the power loss would be:

$R = 2.00\ \Omega$

$I = 2.65 \times 10^6 \text{ A}$

$$P_{\text{loss}} = I^2R = (2.65 \times 10^6)^2(2.00)$$
$$= 1.40 \times 10^{13} \text{ W}$$

This is actually much greater than the power supplied, so 99.995% of the power is lost in transmission!

ii

$P_{\text{Perth}} = 635 \times 10^6 \text{ W}$

$V_{\text{Perth}} = 500 \times 10^3 \text{ V}$

$$I = \frac{P_{\text{Perth}}}{V_{\text{Perth}}} = \frac{635 \times 10^6}{500 \times 10^3}$$
$$= 1.27 \times 10^3 \text{ A}$$

Thus the power loss would be:

$R = 2.00\ \Omega$

$I = 1.27 \times 10^3 \text{ A}$

$$P_{\text{loss}} = I^2R = (1.27 \times 10^3)^2(2.00)$$
$$= 3.23 \times 10^6 \text{ W}$$

Now 0.505% of the power is lost in transmission!

b To determine the voltage at the supply end we need to calculate the voltage drop along the transmission line:

$I = 2.65 \times 10^6 \text{ A}$

$R = 2.00\ \Omega$

$$\Delta V_{\text{Muja-Perth}} = IR = (2.65 \times 10^6)(2.00)$$
$$\Delta V_{\text{Muja-Perth}} = 5.29 \times 10^6 \text{ V}$$
$$V_{\text{Muja}} = \Delta V_{\text{Muja-Perth}} + V_{\text{Perth}}$$
$$= 5.29 \times 10^6 + 240$$
$$= 5.29 \times 10^6 \text{ V}$$

This is 5.29 million volts—totally out of the question of course!

$I = 1.27 \times 10^3 \text{ A}$

$R = 2.00\ \Omega$

$$\Delta V_{\text{Muja-Perth}} = IR = (1.27 \times 10^3)(2.00)$$
$$\Delta V_{\text{Muja-Perth}} = 2.54 \times 10^3 \text{ V}$$
$$V_{\text{Muja}} = \Delta V_{\text{Muja-Perth}} + V_{\text{Perth}}$$
$$= 2.54 \times 10^3 + 500 \times 10^3$$
$$= 5.03 \times 10^5 \text{ V}$$

This is 503 000 volts—much more achievable.

c As the power loss is directly proportional to the resistance, the power loss (at 500 kV) would halve to 1.61×10^6 W. The voltage drop would also halve and so the supply voltage would be 1270 V less. These are not particularly large differences and as the cost of the power line would increase considerably (twice as much metal, and thus stronger towers needed) it would probably not be worthwhile.

You might like to show for yourself that the diameter of a 1.00 Ω aluminium conductor 200 km long needed to carry the power in this example would be about 8.2 cm. You need to know that the resistivity, ρ, of aluminium is $2.65 \times 10^{-8}\ \Omega$ m and that $R = \rho L/A$.

Physics in action — Electrical safety

In the last few decades the use of electrical appliances has boomed, while deaths from electrocution have dropped considerably. One reason for this is the use of earth leakage detectors, or *residual current devices* (RCDs). Any appliance used around the home is connected between the active (±340 V) and neutral (0 V) terminals. Normally, an equal amount of current flows in both conductors. If a fault, or carelessness, allows some current to flow to earth through a person, the active and neutral currents are no longer equal. The RCD is designed to detect this and instantaneously shut off the circuit.

The principle is simple: if the currents in each wire are equal but in opposite directions, their magnetic fields will cancel each other. If the currents are not equal, there will be a net magnetic field. In Figure 4.28, the active and neutral conductors pass through a toroidal core. If the currents are equal, no net field is created. If the currents are not equal, the resultant field will induce a current in the pick-up winding. This current activates a solenoid, which switches off the main supply.

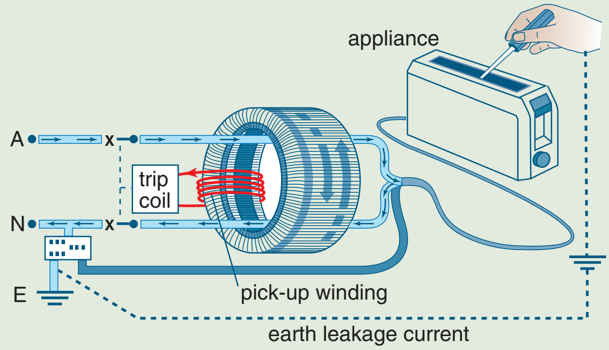


Figure 4.28
Residual current devices prevent fatal shocks by shutting off the current within about 0.03 s.

An RCD cannot eliminate all risk. If you put one hand on a live terminal in the appliance and the other on a neutral wire, the RCD will do nothing because the current going through you is returning through the neutral. For this reason, it is very good practice never to put two hands near a suspect electrical appliance. It is even better practice to disconnect it and take it to an expert!

Electric power: The large scale

Starting up a coal-fired power station takes time. The boilers have to be fired, steam pressure built up and the generators run up to the correct speed of 50 revolutions per second. This can take over 12 hours. It is an equally time-consuming process to shut down again. Unfortunately, we do not use electrical energy at a constant rate. As we wake up, turn on the lights and heater, cook breakfast and then catch the train to school or work, the demand for electricity increases rapidly; typically it might increase by 20% in an hour. Figure 4.29 shows typical variations in the electric power demand in spring, summer and winter.

To cope with this, power companies use the coal-fired stations to supply the 'base load' demand (darker colour) but diesel and gas powered generators for rapid response (which still take some time to heat up) for the intermediate fluctuations in load.

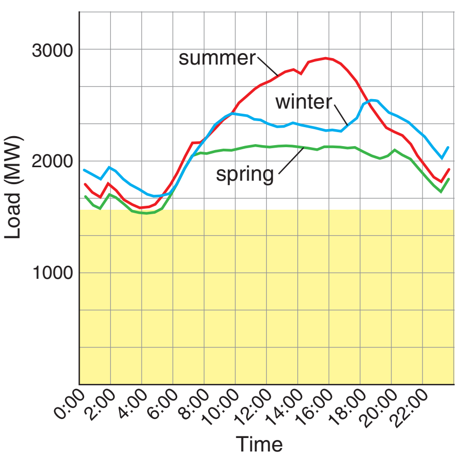


Figure 4.29
Typical load curves for Western Australia on a spring, summer and winter day. Each square represents 667 MW h of energy.

✓ Worked Example 4.5B

a How much electrical energy was used on the summer day shown in Figure 5.29?
b If the power station works at an overall efficiency of 25%, how much coal would have been burnt that day? How much CO_2 would that generate? (The heat value of thermal coal is 22.1 GJ per tonne and each tonne of thermal coal burnt releases 3.35 tonnes of CO_2.)

Solution

a The total energy used is represented by the area under the graph. On this graph, each square represents 333 MW × 2 h, or 667 × 10^3 kW h, which converts to:

$$E_J = E_{kW\,h} \times (3600 \times 1000)$$
$$= (667 \times 10^3)(3.6 \times 10^6)$$
$$= 2.40 \times 10^{12} \text{ J}$$

The number of squares under the summer graph is approximately 80, so the total electrical energy used that day was:

$$E_{J\,total} = (80)(2.40 \times 10^{12})$$
$$= 1.92 \times 10^{14} \text{ J}$$

192 000 GJ, a very large amount of energy!

b At an efficiency of 25%, the total amount of coal energy required is:

$$E_{J\,coal} = \frac{100}{25}(1.92 \times 10^{14})$$
$$= 7.68 \times 10^{14} \text{ J}$$

Each tonne of thermal coal provides 22.1 GJ of energy, so the mass of coal required is:

$$m_{coal} = \frac{7.68 \times 10^{14}}{22.1 \times 10^9}$$
$$= 3.48 \times 10^4 \text{ tonnes}$$

17 300 tonnes of thermal coal is required per day. This would fill about 13 Olympic-sized swimming pools. The mass of carbon dioxide released from burning this coal would be:

$$m_{CO_2} = (3.18 \times 10^4)(3.35)$$
$$= 1.16 \times 10^5 \text{ tonnes of } CO_2$$

This is equivalent to over 2 billion 'black balloons' (50 g of CO_2).

Electric power: The small scale

Power companies measure the electrical energy we use at home by installing a *watt hour meter*. Most are rather like a combination of an AC electric motor and a car odometer: the current being used in the house creates a changing magnetic field, which drives a rotor. The rotor is connected to a series of dials that register the number of kilowatt hours. One kilowatt hour (kW h) is simply the amount of energy used by a 1 kW device in 1 h. For example, an appliance, like a room heater, that uses 2 kW for 5 h will use 10 kW h. Power companies charge around 15 cents for 1 kW h, which is equal to 3.6 MJ (see adjacent Physics file).

The 'electromechanical' watt hour meters described above are gradually being replaced by new electronic 'smart meters', which use semiconductor technology to measure the power used. They also enable functions such as remote reading, charging for power at variable rates and measuring power fed into the grid from household solar or wind generators.

Interactive tutorial

Kilowatt hours

Physics file

The kilowatt hour is simply another energy unit. Energy is the product of power and time:

$$E = P\Delta t$$

1 joule = 1 watt × 1 second
1 kilowatt hour = 1 kilowatt × 1 hour
To convert kilowatt hours to joules:

$$1 \text{ kW h} = 1000 \text{ W} \times 3600 \text{ s}$$
$$1 \text{ kW h} = 3.6 \times 10^6 \text{ J}$$
$$1 \text{ kW h} = 3.6 \text{ MJ}$$

or

$$1 \text{ MW h} = 3.6 \text{ GJ}$$

The future

We do not need to be reminded that modern technological societies consume vast amounts of energy of all sorts. Per head of population, Australians consume around eight times the world average and over 50 times that in developing countries. Roughly one-third of our energy consumption is electrical energy; the rest is mainly oil for transport as well as oil, gas and coal used directly for agriculture, manufacturing and industry.

Electricity itself is a very clean and adaptable form of energy, but it must be generated from some form of primary energy. Virtually any source of energy can be used to create electricity, mostly by producing steam to run a turbine and generator. In Australia, about 77% of our electrical energy comes from coal and 16% from oil and gas. The other 7% is mainly hydro-electricity, with only a very small contribution (about 0.5%) from wind, solar and other renewable sources. In many technologically developed countries, a considerable proportion of electrical energy is generated by nuclear power, for example 34% in the European Union and 28% in Japan. Australia has the largest (24%) known reserves of uranium in the world and currently supplies about 20% of the world's uranium for nuclear fuel.

Figure 4.30
Modern semiconductor technology can efficiently convert the DC voltage from photovoltaic cells into AC voltage and feed it into the grid. New nanotechnology will reduce the cost considerably.

The great challenge for the 21st century is to produce electrical energy without the emission of damaging amounts of the greenhouse gas, carbon dioxide.

While the potential for further hydroelectricity is virtually exhausted, Australia receives vast amounts of solar energy—in the form of direct sunlight and wind. Although at present the cost of these sustainable forms of energy is relatively high, once the real cost of carbon emissions is taken into account there will be a much greater incentive to develop these clean forms of power. It is hoped that geothermal energy ('hot rocks') can also be developed in various parts of the country.

Physics in action — High-voltage DC power transmission and the 'base load' question

The limit on the voltage of an electrical power transmission line is the corona effect—the point at which the intense electric field around the cable breaks down the insulating properties of the air. This limit is around 800 kV, which is a little over the peak voltage reached by a 500 kV (RMS) power line. However, the voltage on an AC line only reaches this briefly twice each cycle and so much of the cycle is 'wasted' in terms of carrying current. On the other hand, if a steady DC voltage of 800 kV was used, the effective current, and hence power carried, would be considerably greater—over twice as much in fact.

The big advantage of AC transmission over DC transmission is that AC voltages can easily be changed by transformers. This is particularly important for distributing power around a city where only the last kilometre or so can be at 240 V. In order to change the voltage of DC, it has been necessary to convert it to AC (using an *invertor*), put it through a transformer and then convert it back to DC (with a *rectifier*). However, because of developments in semiconductor technology, which have made high-current, high-voltage thyristors (a type of transistor) available, it is now possible to change high-voltage AC into high-voltage DC and vice versa with very little loss of power. This means that it has become practicable to use high-voltage DC (HVDC) transmission lines.

There are several advantages in using HVDC over AC for long transmission lines. The first was mentioned above—more than twice the power can be sent along the same size cables. Second, any power line has capacitance, which means some current is 'lost' in charging up the line. This happens every cycle for AC, but not at all for DC. There are significant losses associated with this once a power line is more than a few hundred kilometres long. Third, because of the varying current in an AC line, it is surrounded by a varying magnetic field, which we know will induce currents in any conductor nearby. This is why high-voltage AC power lines need to be so far above the ground and why they can't be put underground or under the sea for any distance. It also means power is lost by electromagnetic radiation from the lines. All these factors limit AC transmission lines to a few hundred kilometres.

Figure 4.31

At the heart of an AC-DC converter station are valves comprising power thyristors, which are basically large high voltage transistors.

While the cost of the inverting and rectifying equipment is relatively high, because HVDC transmission does not suffer the problems of AC, it becomes more economical for large distances—up to thousands of kilometres. In fact, losses of only 3% per 1000 km have been achieved.

Another advantage of HVDC transmission is that different power systems can be linked. Any generator linked to an AC grid must exactly match the frequency and phase—they have to be 'synchronous'. This limits the size of a grid. However, unsynchronised grids can be linked by HVDC. The invertor which connects the DC to the new AC system synchronises its output to the new grid. This is indeed how the Victorian and Tasmanian power systems have been linked together by the 400 kV Basslink cable, which consists of one high-voltage undersea cable about 12 cm in diameter and a 9 cm return cable in the same trench. It can transfer over 500 MW of power in either direction between the two states.

It is often said that sustainable energy systems such as wind and solar cannot supply 'base load' electric power. But with the possibility of linking distant systems, for example geothermal from outback Western Australia, wind from the south-west and from Geraldton and solar from Perth rooftops, the need for large baseload stations is reduced considerably.

4.5 SUMMARY Distributing electricity

- The power delivered by an electric transmission line is equal to the product of the current and voltage. High power requires high current and/or voltage.
- The power lost whenever current flows through transmission lines is equal to I^2R or $\frac{V^2}{R}$.
- Because the power loss is proportional to the square of the current, it is important to reduce the current in long-distance transmission lines by using very high voltages.
- The practical upper limit to the transmission voltage is around 500 kV.
- The demand for electric power varies during the day. The base load is normally provided by coal-fired plants and the peak demand by diesel or gas turbines.
- Total energy demand is represented by the area under a load–time graph ($E = P\Delta t$).
- The kilowatt hour (kW h) is a useful unit for electric energy. It is equal to 3.6 MJ.

4.5 Questions

All voltages referred to in these questions are RMS.

1 In Western Australia, electric power is transmitted at very high voltages (up to 500 kV).
- **a** What is the main reason for this?
- **b** What factors limit the use of even higher voltages for power transmission?

2 A power station generates 500 MW of electrical power which is fed to a transmission line. What current would flow in the transmission line if the input voltage is:
- **a** 250 kV?
- **b** 500 kV?

3 A 100 km transmission line made from aluminium cable has an effective radius of 1.00 cm and a total resistance of 10.0 Ω. The line carries the electrical power from the 500 MW power station to a substation. Calculate the percentage power loss in the line when the power station is operating at:
- **a** 250 kV
- **b** 500 kV.

4 A transmission line made from aluminium cable is twice the length (200 km) and twice the radius (2.00 cm) of the line described in Question 3. How will the percentage power loss in this line compare to that in the line in Question 3 when it is operating at 500 kV?

The following information applies to questions 5 and 6.

A solar-powered generator produces 5.00 kW of electrical power at 500 V. This power is transmitted to a distant house via twin cables of total resistance 4.00 Ω.

5
- **a** What is the current in the cables?
- **b** What is the total power loss in the cables?
- **c** What is the percentage power loss in the cables?
- **d** What is the voltage available at the house?

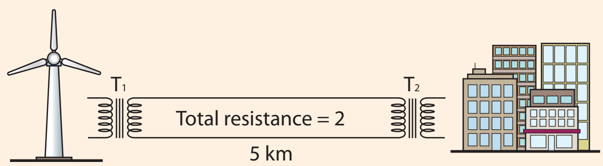
6 The supplier is unhappy with this power loss and uses a transformer to step up the transmission voltage to 5.00 kV, and a step-down transformer at the other end. The power output from the generator remains at 5.00 kW and the same 4.00 Ω cables are used.
- **a** What current now flows in the cables?
- **b** Calculate the percentage power loss for this new arrangement.
- **c** What is the voltage output to the house?

7 If the cost of electricity in Western Australia is 14.0 cents per kilowatt hour, determine the cost of running the following.
- **a** 1.00 kW heater for 2.00 hours
- **b** 80.0 W light globe for 30.0 minutes
- **c** 250.0 W television for 12.0 hours
- **d** 6.00 W clock radio for a week
- **e** computer printer using 3.00 W on standby for a year

8 A town 100.0 km from a power station uses up to 500.0 MW of power. The power lines between the power station and the town have a total resistance of 2.00 Ω.
- **a** If this power was to be transmitted at 250.0 V, how much current would be required and what would the voltage loss along the power line be? Would this be practical?
- **b** If the power from the generator is transmitted at 100.0 kV, what current is required and what would be the voltage drop along the line? What is the voltage at the town?
- **c** If another power line is added so that the total resistance halves, to 1.00 Ω, how much power would be lost in the lines?

The following information applies to questions 9 and 10.

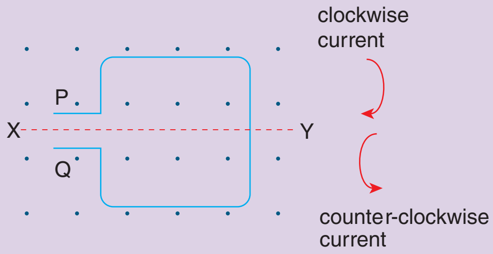
The diagram shows a wind turbine which runs a 150.0 kW generator with an output voltage of 1000 V. The voltage is increased by transformer T_1 to 10 000 V for transmission to a town 5.00 km away through power lines with a total resistance of 2.00 Ω. Another transformer, T_2, at the town reduces the voltage to 250.0 V. Assume that the transformers are 'ideal'.



9 When the system is running at full power:
- **a** what is the current in the power line?
- **b** what is the voltage drop along the power line and the voltage at the input to the town transformer?
- **c** how much power is lost in the power line? Is this a problem?

10 It is suggested that the cost of the first transformer could be avoided if the generator was connected to the power line directly and T_2 reduced the voltage at the end of the line to 250.0 V for the town.
- **a** What current would now flow through the transmission lines?
- **b** What voltage drop would there be along the transmission lines and what is the input voltage to T_2?
- **c** What power would be lost in the lines this time and how much power is available to the town?
- **d** Was it a good idea to use this scheme?

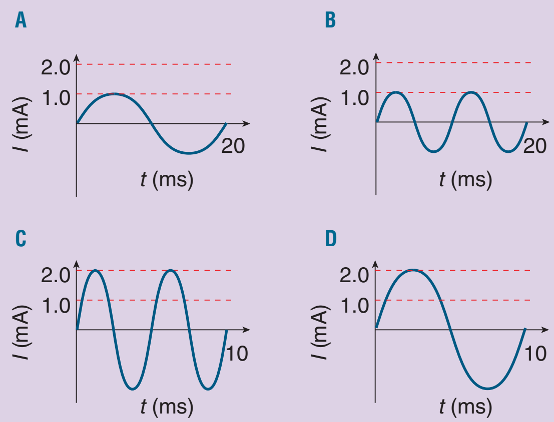
The following information applies to questions 1–3.
A rectangular coil of area 40.0 cm^2 and resistance 1.00 Ω is located in a uniform magnetic field $\boldsymbol{B}_{\perp} = 8.00 \times 10^{-4}$ T which is directed out of the page. The plane of the coil is initially perpendicular to the field as depicted in the diagram below.



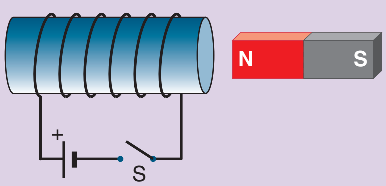
1 State the magnitude and direction of the average current induced in the coil when the following changes are made.
  a The magnitude of B is doubled in 1.00 ms.
  b The direction of B is reversed in 2.00 ms.
  c The magnitude of B is halved in 1.00 ms.

2 The coil is rotated about the axis XY with constant frequency of $f = 100$ Hz.
  a What is the maximum EMF induced in the coil?
  b What is the peak current induced in the coil?

3 Which one of the graphs A–D best describes the current–time relationship for the coil for the following values of frequency and field strength?
  a $f = 50.0$ Hz, $B = 8.00 \times 10^{-4}$ T
  b $f = 200$ Hz, $B = 4.00 \times 10^{-4}$ T
  c $f = 100$ Hz, $B = 4.00 \times 10^{-4}$ T

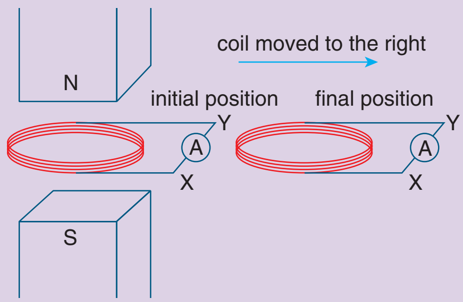


4 An electromagnet with a soft iron core is set up as shown in the diagram below. A small bar magnet with its north end towards the electromagnet is placed to the right of it. The switch S is initially open. The following questions refer to the force between the electromagnet and the bar magnet under different conditions.



  a Describe the force on the bar magnet while the switch remains open.
  b Describe the force on the bar magnet when the switch is closed and a heavy current flows.
  c The battery is removed and then replaced so that the current flows in the opposite direction. Describe the force on the bar magnet now.

The following information applies to questions 5–7.
During a physics experiment a student pulls a horizontal coil from between the poles of two magnets in 0.100 s. The initial position of the coil is entirely in the field while the final position is free of the field. The coil has 40 turns, each of radius 4.00 cm, and a total resistance of 2.00 Ω. The field strength between the magnets is 20.0 mT.

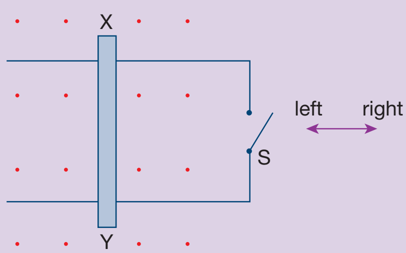


5 What is the magnitude and direction of the average current induced in the coil as it is moved from its initial position to its final position?

6 The student makes some modifications to the equipment and then repeats the experiment. Which one or more of the following would result in a greater induced EMF in the coil than for the original situation?
  A The number of turns in the coil is increased.
  B The area of each turn is reduced.
  C The coil is pulled out of the field faster than before.
  D The direction of the magnetic field is reversed before the coil is pulled out.

7 What would be the direction of the induced current in the coil when the student moved it from its final position back to its original position? Justify your answer.

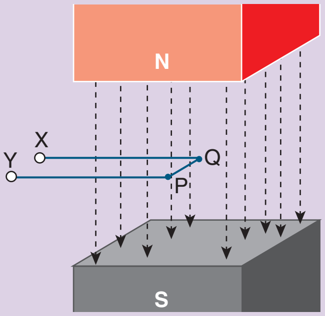
The following information applies to questions 8 and 9.
A copper rod, XY, of length 20.0 cm is free to move along a set of parallel conducting rails as shown in the following diagram. These rails are connected to a switch S, which completes a circuit with a total resistance 1.00 Ω when it is closed. A uniform magnetic field of strength 10.0 mT, directed out of the page, is established perpendicular to the circuit. S is closed and the rod is moved to the right with a constant speed of 2.00 m s^{-1}.



8 **a** What is the magnitude and direction of the average current induced in the rod?
  b What is the magnitude and direction of the magnetic force on the rod due to the induced current?

9 The force moving the rod is removed and the rod is stationary. The switch is now opened. What is the direction of the induced current in the rod? Justify your answer.

The following information applies to questions 10–13.
The following diagram shows a section of a conducting loop XQPY, part of which is placed between the poles of a magnet whose field strength is 1.00 T. The side PQ has length 5.00 cm. X is connected to the positive terminal of a battery while Y is connected to the negative terminal. A current of 1.00 A then flows through this loop.



10 What is the magnitude of the force on side PQ?

11 What is the direction of the force on side PQ?

12 What is the magnitude of the force on a 1.00 cm section of side XQ that is located in the magnetic field?

13 The direction of the current through the loop is reversed by connecting X to the negative terminal and Y to the positive terminal of the battery. What is the direction of the force on side PQ?

The following information applies to questions 14 and 15.

A ship with a vertical steel mast of length 8.00 m is travelling due west at 4.00 m s^{-1} in a region where the Earth's magnetic field (assumed horizontal) is equal to 5.00×10^{-5} T north.

14 **a** What average EMF would be induced between the ends of the mast?

b A crew member connects a rectangular wire frame across the ends of the mast, forming a 40.0 m^2 loop whose plane is perpendicular to the Earth's field. The circuit formed by the loop and the mast has a total resistance of 8.00 Ω. What would be the average current induced in the circuit? Justify your answer.

15 The ship is still moving with the same velocity but now encounters a mysterious region where the Earth's magnetic field is changing at a rate of 1.00×10^{-5} T s^{-1}. Calculate the average current induced in the circuit previously described.

16 A rectangular conducting loop of dimensions 100.0 mm × 50.0 mm and resistance $R = 2.00$ Ω is located with its plane perpendicular to a uniform magnetic field of strength $B = 1.00$ mT.

a Calculate the magnitude of the magnetic flux Φ_B threading the loop.

b The loop is rotated through an angle of 90.0° about an axis, so that its plane is now parallel to $\boldsymbol{B}$. Determine the magnetic flux Φ_B threading the loop in the new position.

c The time interval for the rotation $\Delta t = 2.00$ ms. Determine the average EMF induced in the loop.

d Determine the value of the average current induced in the loop during the rotation.

e Will the current keep flowing once the rotation is complete and the loop is stationary? Explain your answer.

17 A 5.000 Ω coil, of 100 turns and radius 3.00 cm, is placed between the poles of a magnet so that the flux is a maximum through its area. The coil is connected to a sensitive current meter that has an internal resistance of 595 Ω. It is then moved out of the field of the magnet and it is found that an average current of 50.0 μA flows for 2.00 s.

a Had the coil been moved out more quickly so that it was removed in only 0.5 s, what would have been the average current?

b What is the strength of the magnetic field?

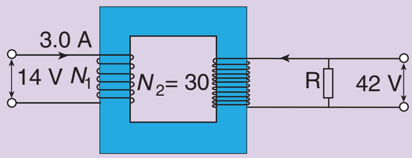
18 A physics student constructs a simple generator consisting of a coil of 400 turns. The coil is mounted on an axis perpendicular to a uniform magnetic flux density $\boldsymbol{B}_{\perp} = 50.0$ mT and rotated at a frequency $f = 100$ Hz. It is found that during the rotation, the peak voltage produced is 0.900 V.

a Sketch a graph showing the voltage output of the generator for at least two full rotations of the coil. Include a scale on the time axis.

b What is the RMS voltage generated?

c The student now rotates the coil with a frequency $f = 200$ Hz. How would your answers to parts a and b be affected?

19 An ideal transformer is operating with an RMS input voltage of 14.0 V and RMS primary current of 3.00 A. The output voltage is 42.0 V. There are 30 turns in the secondary winding.

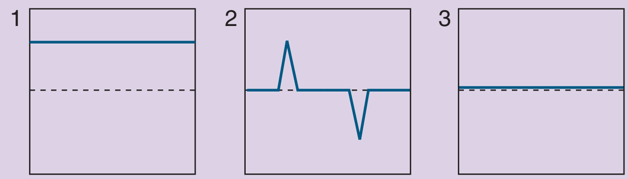


a What is the RMS output current?

b How many turns are there in the primary winding?

c How much power is consumed by the resistor R?

20 The following diagrams show the output voltages for a transformer as they appear on the screen of a CRO. The broken line in each display is the time base and represents zero voltage.

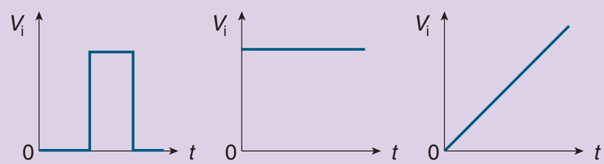


Which of the following input voltages A–C would produce:

a display 1?

b display 2?

c display 3?



21 Which of the following is the best description of how a transformer transfers electrical energy from the primary windings to the secondary windings?

A The current through the primary windings produces a constant electric field in the secondary windings.

B The current through the primary windings produces a steady magnetic field in the secondary windings.

C The current through the primary windings produces a changing magnetic field in the secondary windings.

22 When a transformer is plugged in to the 240 V mains but nothing is connected to the secondary coil, very little power is used. The best explanation for this is that:

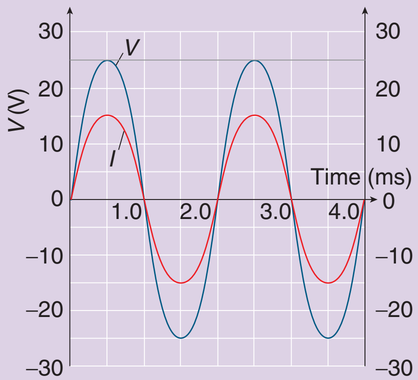
A The primary and secondary coils are in series and so no current can flow in either if the secondary coil is open.

B There can be no magnetic flux generated in the transformer if the secondary coil has no current in it.

C The magnetic flux generated by the current in the primary produces an EMF that opposes the applied voltage.

D The magnetic flux generated by the secondary coil almost balances out that due to the primary coil.

The following information applies to questions 23–25.
A physics student uses a CRO to display the current, I, through and the potential difference, ΔV, across the terminals of a loudspeaker which has been connected to a signal generator. The graph obtained from the CRO display is shown below.



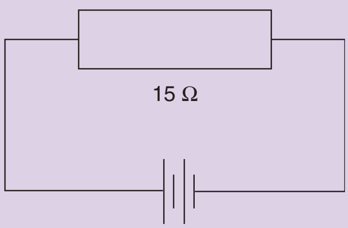
23 **a** What is the frequency of the output signal from the generator?
b What is the RMS voltage of this signal?
c What is the peak-to-peak voltage for the signal?

24 **a** Calculate the RMS power output of the signal generator.
b What is the peak power output of the generator?
c Calculate the apparent resistance of the speaker.

25 The student decides to test the power output of a new stereo amplifier. The maximum RMS power output guaranteed by the manufacturer (assumed accurate) is 60.0 W. Which set of specifications is consistent with this power output?

	Peak–peak voltage (V)	Peak–peak current (A)
A	20.0	3.00
B	40.0	6.00
C	40.0	12.00
D	20.0	6.00

26 Which set of specifications in Question 25 would best describe the AC voltage–current combination that would produce a power consumption equivalent to the DC circuit shown in the following diagram?



27 A student builds a simple alternator consisting of a coil containing 500 turns, each of area 10.0 cm^2, mounted on an axis that can rotate between the poles of a permanent magnet of strength 80.0 mT.
At a frequency of 50.0 Hz, it is found that the peak voltage produced is 12.6 V.
a What are the peak-to-peak and RMS voltages?
b If the frequency is doubled to 100.0 Hz, how will the peak and RMS voltages change?
c What frequency of rotation is required for the alternator to produce a voltage of 16.0 V RMS?
d If the magnetic field is reduced to 60.0 mT, what will be the peak voltage at 50.0 Hz?

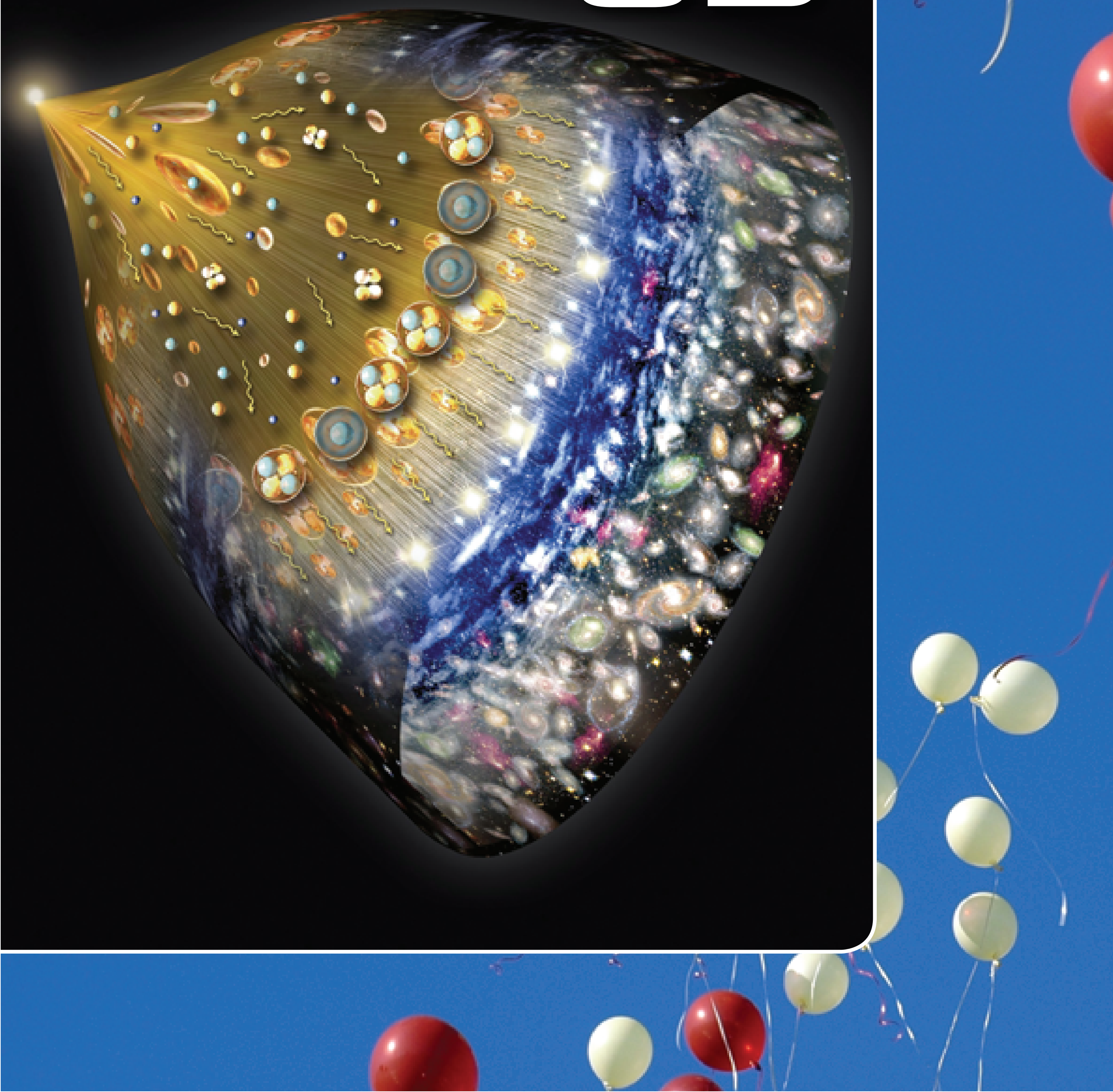
28 A farmer needs to operate a 24.0 V, 480.0 W machine that is a long way from the nearest 240.0 V power supply. He has bought an 'ideal' transformer with a turns ratio of 10 : 1. The generator is to be connected to the machine by a long twin-core cable with a total resistance of 2.00 Ω. The farmer initially uses the cable to supply the 240.0 V to the transformer, which is near the machine.
a How much current would be flowing in the cable, and what would be the voltage at the input to the transformer? Will the machine operate satisfactorily?
b He is concerned that the cable is carrying a dangerous voltage and so decides to put the transformer at the power supply end. When he connects the machine at the other end, he finds that it does not operate properly. Explain why this would be the case.
He then buys a different transformer that he calculates will operate the machine correctly at 24 V and 20 A.
c What transformer output voltage will be needed, and what is the turns ratio of the new transformer?
d How much power is being wasted in this arrangement?

The following information applies to questions 29 and 30.
A generator is to be installed in a farm shed to provide 240.0 V power for the farmhouse. A twin conductor power line with total resistance 8.00 Ω already exists between the shed and house. The maximum power requirement of the house is estimated as 2000 W. The farmer has seen a cheap 240.0 V DC generator advertised and is tempted to buy it.

29 What would you advise the farmer about using the 240.0 V DC generator? Consider as many factors as you can. Why is he likely to damage some of his appliances while some would work normally?

30 The farmer is now convinced that he needs an AC generator and that he wants no more than a 5.00 V drop in voltage at full power usage. He realises that he will have to buy a higher voltage generator and use a transformer at his house. He finds a generator that produces 1200 V AC. What sort of transformer should he buy? What voltage would he find in the house with very little load, half load and full load? Would this set-up suit his purpose?

Physics 3B



The following context section and the content chapter Working in Physics on the CD, along with Chapters 5–7, cover the content required for the Physics course 3B.

Outcomes

The Physics Outcomes are as follows.

Outcome 1: Investigating and communicating in physics

Students investigate physical phenomena and systems, collect and evaluate data, and communicate their findings.
In achieving this outcome, students:

- develop questions and ideas about the physical world to prepare an investigation plan
- conduct experiments and investigations
- analyse data and draw conclusions based on evidence
- evaluate the accuracy and precision of experimental data and the effectiveness of their experimental design
- communicate and apply physics skills and understandings in a range of contexts.

Outcome 2: Energy

Students apply understanding of energy to explain and predict physical phenomena.
In achieving this outcome, students:

- apply understanding of conceptual models and laws relating to energy
- apply understanding of mathematical models and laws relating to energy.

Outcome 3: Forces and fields

Students apply understanding of forces and fields to explain physical phenomena.
In achieving this outcome, students:

- apply understanding of conceptual models and laws relating to forces and fields
- apply understanding of mathematical models and laws relating to forces and fields
- apply understanding of the vector nature of some physical quantities.

It is envisaged that students will fulfil the requirements of Outcome 1 through investigative work in the classroom. Investigative work will typically be related to the content covered in Chapters 5–7. Working in Physics covers the basic requirements for understanding and processing measured quantities in the investigative work. A complete chapter that revises and expands on Working in Physics for Stage 3 can be found on the *ePhysics* CD.

Chapters 5–7 cover the required content for Outcomes 2 and 3.

Unit description

Unit 3B focuses on two broad areas of physics: ***Particles, Waves and Quanta*** (Chapters 5 and 6) and ***Motion and Forces in Electric and Magnetic Fields*** (Chapter 7).

In ***Particles, Waves and Quanta,*** students extend their understanding about the nature and behaviour of waves, including sound and light. They learn how electromagnetic waves interact with matter, how they are produced and the characteristics of their spectra. They extend their understanding of the fundamental particles of nature to include quarks and leptons. They also learn some modern physics, including Einstein's Special Theory of Relativity, and some fundamentals of cosmology.

In ***Motion and Forces in Electric and Magnetic Fields***, students learn about the concepts of electric fields and how energy can be stored in the field. They also learn about the forces that charged particles experience when moving in electric and magnetic fields.

Contexts

The context material, preceding the content chapters, supports the content of Unit 3B:

- The search for understanding (CD)
- The sounds of music.

The sounds of music

By the end of this context

you will have covered material including:

- understanding musical sound
- calculating power, intensity and level of sound
- exploring the characteristics of a pendulum
- understanding how waves interact in strings and tubes
- recognising the properties and behaviour of waves in one and in two dimensions
- understanding how magnetic forces can be used to create sound.

Mention the name Pythagoras and you will, hopefully, think of a right-angled triangle and 'that theorem'. What is not so well known about the Greek philosopher is that he is credited with being the first person to investigate the mathematical relationships evident in nature. In this case, he discovered the pattern of the ratios of lengths of string and the sound that emanates from them when plucked.

Music is all around us. It is used to stir the emotions, to create atmosphere in movies, and to grab our attention in advertising. Many of us have at least played a recorder and some have gone on to play other instruments. So how do these instruments work? What is the physics behind them? Why do instruments sound different from one another and how are electronic keyboards able to produce the sound of specific instruments? These and many other questions will be answered in 'The sounds of music'.

Figure sm.1 Not many people know that Pythagoras revealed the relationship between the length of a string and the sound that it makes when bowed or plucked.

The elements of music

The sound we hear from instruments is described by musicians using words such as volume (how loud the sound is), pitch (whether the sound is 'high' or 'low') and timbre (pronounced 'tamber'; the characteristic sound of a specific instrument). It is the timbre of sounds from different instruments that allows us to distinguish between the sound of a middle C played on a guitar from that played on a violin.

These are subjective qualities of sound, dependent, in part, on the listener. This means that what one person perceives as being loud, for example, may not be loud to another. The explanation of timbre requires us to look at harmonics but first we must investigate volume and pitch.

Investigation

Investigate the sounds made when you blow across the top of a soft-drink bottle and then do it again with the bottle a quarter full and then half full. In each case, make a mental note of the sound that is heard. What words can we use to describe the sound that was heard and what effects did the amount of water have on this sound?

Volume

Volume, or loudness, is related to the physical quantity of intensity, *I*, defined as the power passing through each metre squared of the surrounding space, as defined by the following equation:

$$I = \frac{P}{A}$$

where I is the intensity (W m^{-2}), P is the power (W) and A is the area (m^2).

The units of intensity are watts per square metre (W m^{-2}) or, in terms of energy, joules per square metre per second (J m^{-2} s^{-1}). The power of a sound wave is directly proportional to the square of the wave's amplitude, $P \propto \text{amplitude}^2$.

See the wave equation on page 190.

The total power of the sound wavefront remains constant as the sound spreads out in a spherical 'bubble'. As the sound wavefront spreads out, the intensity of the sound decreases as the area of the sphere increases:

$$P_1 = P_2$$
$$I_1 A_1 = I_2 A_2$$
$$I_1 4\pi\, r_1^2 = I_2 4\pi\, r_2^2$$
$$I_1\, r_1^2 = I_2\, r_2^2$$

where P is the power of the wave (W) at two locations, 1 and 2, I is the intensity (W m^{-2}) of the sound at 1 and 2, A is the area (m^2) of the sphere of sound ($A = 4\pi r^2$) at 1 and 2 and r is the radius of the sphere of sound from the source (m).

The intensity of sound follows the inverse square law like so many other variables in nature. That is, if you halve the distance you are from a source of music, the intensity will increase by a factor of four. Halve the distance again and the intensity increases to sixteen times the original intensity (Figure sm.2).

✓ Worked Example sm.1

When 3.00 m from a sound source the intensity is 2.50×10^{-6} W m^{-2}. Calculate the intensity from the same source at a distance of 10.0 m.

Solution

$I_1 = 2.50 \times 10^{-6}$ W m^{-2}

$r_1 = 3.00$ m

$r_2 = 10.0$ m

$$I_1 r_1^2 = I_2 r_2^2$$

$$I_2 = \frac{I_1 r_1^2}{r_2^2} = \frac{(2.50 \times 10^{-6})(3.00)^2}{10.0^2}$$

$$= 2.25 \times 10^{-7} \text{ W m}^{-2}$$

Hence, the intensity from the same source at the further distance is 2.25×10^{-7} W m^{-2}.

The range of human hearing ranges from an extremely loud 1.00 W m^{-2} to the softest sound we can hear at 10^{-12} W m^{-2}. This is a huge range in intensities over which we can hear; in fact, the loudest sound is a million million times the quietest sound. It makes more sense to convert this range to a log scale of the ratio of the sound compared to the quietest sound, which becomes the reference intensity ($I_0 = 1 \times 10^{-12}$ W m^{-2}). This converts the sound level scale from one that ranges from 10^0 to 10^{12} to a scale that uses the indices of 10, which range from 0 to 12. This scale for sound level has the units of bel (B) after Alexander Graham Bell. If we multiply this scale by 10 we can measure sound levels in decibels (dB). The equation that calculates the absolute sound level (L) of a sound is:

$$L = 10\log\left(\frac{I_1}{I_0}\right)$$

where L is the sound level (dB), I_1 is the intensity of a sound (W m^{-2}) and I_0 is the reference intensity, which is equal to 1×10^{-12} W m^{-2}.

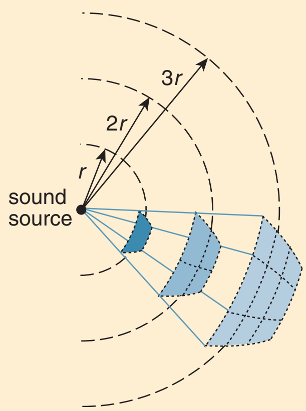


Figure sm.2 As it moves away from the sound source, a sound wave spreads radially. The energy per unit area decreases with the square of the distance from the source, so the intensity of the sound decreases by the same ratio.

A sound level meter is used to measure the decibels of a sound. A sound intensity of 1×10^{-12} W m^{-2} corresponds to 0 dB (refer to Table sm.1), and is considered to be the lowest sound intensity an average person can hear when the frequency of the sound is 1000 Hz.

Sometimes it is handy to be able to compare two intensities, as opposed to comparing one intensity to the reference intensity (I_0). The number of decibel difference between two intensities is defined by the following equation:

$$\Delta L = 10\log\left(\frac{I_2}{I_1}\right)$$

where ΔL is the difference between sound levels (dB), I_1 is the intensity of one sound (W m^{-2}) and I_2 is the intensity of another sound (W m^{-2}).

✓ Worked Example sm.2

a Calculate the sound level of each of the two sounds from Worked Example sm.1 in terms of decibels.
b Compare the two intensities from Worked Example sm.1 in terms of decibels.

Solution

a

$I_1 = 2.50 \times 10^{-6}\ \text{W m}^{-2}$

$I_0 = 1 \times 10^{-12}\ \text{W m}^{-2}$

$$L_1 = 10\log\frac{I_1}{I_0} = 10\log\frac{2.50 \times 10^{-6}}{1 \times 10^{-12}}$$
$$= 10\log(2.50 \times 10^{-6})$$
$$= 64.0\ \text{dB}$$

$I_2 = 2.50 \times 10^{-6}\ \text{W m}^{-2}$

$I_0 = 1 \times 10^{-12}\ \text{W m}^{-2}$

$$L_2 = 10\log\frac{I_2}{I_0} = 10\log\frac{2.25 \times 10^{-7}}{1 \times 10^{-12}}$$
$$= 10\log(2.25 \times 10^{5})$$
$$= 53.5\ \text{dB}$$

b

$I_1 = 2.50 \times 10^{-6}\ \text{W m}^{-2}$

$I_2 = 2.25 \times 10^{-7}\ \text{W m}^{-2}$

$$\Delta L = 10\log\frac{I_2}{I_1} = 10\log\frac{2.25 \times 10^{-7}}{2.50 \times 10^{-6}}$$
$$= 10\log(9.00 \times 10^{-2})$$
$$= 10.5\ \text{dB}$$

Hence, the sound level decreased by 10.5 dB as the listener moved from 3.00 m to 10.0 m from the source.

Table sm.1 gives the decibel readings as would be read on a sound level meter for a variety of sound sources. Also indicated is where the music terms 'piano', *p* (or soft), and 'forte', *f* (or loud), would be.

table sm.1 Level of sound intensity for a range of sounds

Sound	Level of sound intensity (dB)
Threshold of hearing	0
Rustling of leaves	10
Suburban home (*ppp*)	40
Noise inside a large shop (*p*)	60
Talking at 1 m distance	65
Hair dryer (*f*)	80
Inside car in city traffic	90
Car without a muffler (*fff*)	100
Live rock concert	120
Threshold of pain	130

Note that an increase of 10 dB corresponds to an increase in intensity by a factor of ten but that a person will, generally, perceive the sound to be only twice as loud. This is the subjective aspect of the perception of sound. In general, each increase of 10 dB corresponds to a doubling of the perceived loudness. Conversely a decrease of 10 dB corresponds to a perceived halving of the loudness.

Investigation

Carry out some research to extend Table sm.1 to include appliances that emit low level but very persistent noise, such as air-conditioners and computers. Also investigate the long-term effects of such exposure.

Pitch

The pitch of a sound is how high (like a whistle) or low (like the rumble of a passing truck) it is. Pitch is associated with the physical quantity of frequency.

A child may have perfect hearing and be able to detect longitudinal waves passing through air that have a frequency between 20 Hz and 20 kHz. These waves would be sound waves to the child. Frequencies below 20 Hz are termed infrasonic or subsonic and those above 20 kHz are termed ultrasonic.

Very few natural sources of sound produce monotones, or single frequency sounds. We are constantly bombarded with a vast range of frequencies, either from a variety of sources or from a single source such as a musical instrument. Music is created by instruments that produce frequencies of up to about 4000 Hz, but they produce a range of frequencies while producing a single note. This is what produces the timbre of an instrument. Figure sm.3 indicates the range of frequencies generated by a variety of instruments.

Practical activity

15 Pitch, loudness and quality

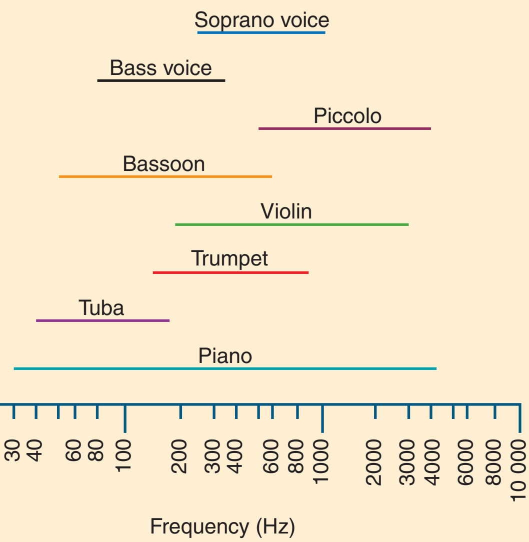


Figure sm.3 The range of frequencies generated by a variety of instruments.

Investigation

Sound-system amplifiers have bass and treble knobs. Investigate the effect these knobs have on the music heard from some source (e.g. CD, DVD, cassette tape, television). What range of frequencies do the knobs influence?

Extended response

1. What is a graphic equaliser and what is its purpose?
2. Research the hearing sensitivity of other animals. How do they use sound as a means of communication?
3. What are some causes of deafness and how does our hearing change with age?

In terms of perception, loudness and pitch are related. Our ears are more sensitive to certain ranges of frequencies, as shown in Figure sm.4. We are generally more sensitive to frequencies between 50 Hz and 9 kHz. Our greatest sensitivity is for frequencies around 3 kHz.

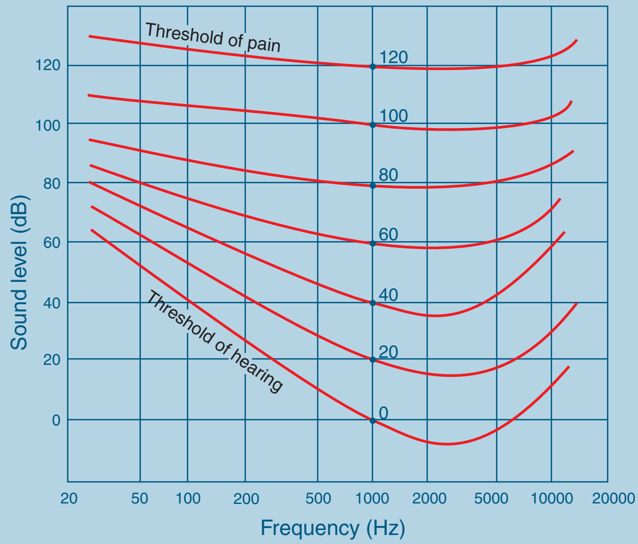


Figure sm.4 Graph of the frequencies of sound waves that are audible to the human ear. The large dots represent the reference loudness, which is the perceived loudness of a 1000 Hz sound at different sound levels. The curves indicate the sound levels that produce the same perceived loudness at other frequencies. For example, a sound of frequency 50 Hz must have a sound level of 80 dB to be perceived to be as loud as a 1000 Hz sound at 60 dB or a 10 000 Hz sound of 70 dB.

Extended response

1. Use available resources to find out what Pythagoras discovered about musical instruments and the sounds they make.
2. Find out the expression for power transmitted by a transverse wave through a guitar string.

Exercises

E1 A physics teacher asks his class to work quietly, but records a sound level of 42.0 dB. What is the intensity (in W m^{-2}) of the class?

E2 A physics student at the school ball records the sound level in front of the speakers using a sound meter. She records the intensity of the sound to be 98.0 dB at 3.00 m. Determine the intensity if she takes two steps back from the speaker, at 5.00 m.

Generating sound

All sound is a consequence of pressure waves passing through the air after being created by an oscillating body. Music is created when these oscillations are controlled and chosen by the playing of instruments.

All instruments work on the following three processes:

1 The **generator** initiates a vibration.
2 The **resonator** vibrates at the 'right' or resonant frequency.
3 The **radiator** projects the vibrations into the surrounding air.

Instruments are generally classified into one of four groups: strings, brass, woodwind and percussion. It is the nature of the generator, resonator and radiator that dictates the classification of an instrument.

Investigation

Either with students in your class or by inviting musicians in, explore the different sounds created by a range of instruments. What relationships are there between the design of various instruments and the sound they produce? Use a CRO (cathode ray oscilloscope) and microphone to view the waveforms produced by the instruments.

Activity

Practical activity: The plucked string

Purpose

To investigate the factors affecting the sound produced by a plucked string.

Materials

- a monochord (a horizontal string held taut by a variable mass hung over a pulley)
- a collection of brass masses, preferably of known mass
- strings of different thicknesses
- CRO (either a traditional one or a computer software version)

Procedure

1 Attach one end of a string to something solid like a piece of wood. Attach the other end of the string to a device for attaching the brass masses. Hang the latter end of the string over a pulley at the edge of a bench. Pluck the string and listen to the sound and observe the waveform produced on the CRO. Vary the length by sliding the wood towards or away from the edge of the bench. Again listen and observe the waveform on the CRO.
2 Using an appropriate time scale on the CRO, determine the frequency of the sound wave for each string length tested.
3 Vary the tension in the string by changing the number of brass masses hanging from the string, and record the sound wave produced using a CRO.
4 Determine the frequency of the sound wave for each mass used.
5 Repeat steps 1–4 with strings of different thicknesses.

Discussion

What qualitative observations can you make? That is, how does the length, mass and thickness of the string affect the sound produced, in terms of both what is heard and what is observed on the CRO?

Extension

If you have the means, investigate these relationships quantitatively. This implies that you will need to make accurate measurements of lengths and masses and to quantify the thickness of the string. As you will see later in this context, it would be useful to calculate the 'mass per unit length' of the strings used.

String instruments

Einstein (who played the violin) said that the physics of the design and construction of the violin was beyond quantitative analysis. All instruments have evolved rather than been designed. Craftsmen have collectively experimented over hundreds of years to come up with the violin we have today. Having said this, however, there are basic physics concepts that can be investigated that explain how the violin achieves its distinctive sound.

The generator for a violin is the bow. The bow uses lengths of horsehair because scales on the hairs all face in the same direction. The hairs are placed on the stick of the bow so that half of the scales face one way and half, the other. It is these scales that 'catch' the string of the violin and pull it sideways. Once static friction is overcome the string springs back and starts to vibrate; hence the string acts as the resonator. Note that when the string comes back it doesn't get caught by the opposing scales because kinetic friction is less than static friction.

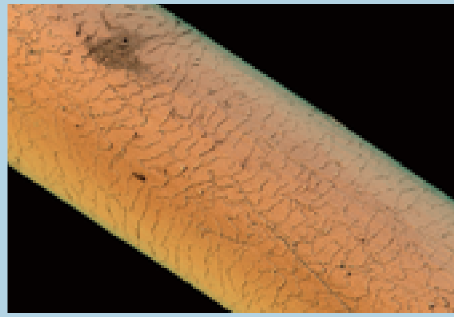


Figure sm.5
Micrograph of a horsehair with scales all facing the same direction.

See 5.2 Wave behaviour on page 192.
See 5.3 Wave interactions on page 201.

Diagrams of standing transverse waves are usually ideal representations. The mode of vibration of the strings of a string instrument is actually far more complicated.

What can be observed, or experienced by violin players, is that the closer the string is bowed to the bridge, the louder the sound, and the greater the number of higher harmonics that can be heard. Violin players can change the frequency at which the strings vibrate by effectively changing their length by pressing the string against the fingerboard (Figure sm.6).

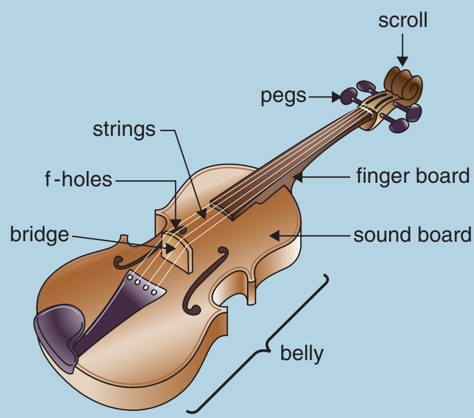


Figure sm.6 The parts of a violin.

The belly of the body, especially the top surface, resonates with the vibrations brought to it by the bridge. The soundboard, or top surface of the belly, acts as the primary radiator. Its job is to take all of the vibrations of the string and to amplify and project them towards the audience. Also, the air within the violin vibrates and its energy is dissipated through the *f*-shaped holes. Bodies made by great instrument makers, such as Antonio Stradivari around 1670, today cost millions of dollars.

Marin Mersenne (1588–1648), a Catholic priest who spent his life in a Paris monastery, is credited with having discovered the laws of stretched strings. He discovered that the frequency of vibration of a stretched string is:

- inversely proportional to its length, $f \propto \frac{1}{L}$
- directly proportional to the square root of the tension in the string, $f \propto \sqrt{T}$
- inversely proportional to the square root of the mass per unit length, $f \propto \frac{1}{\sqrt{\mu}}$.

The following equation expresses the relationships as a single equation.

$$f = \frac{1}{2L}\sqrt{\frac{T}{\mu}}$$

where f is the frequency of vibration in a stretched string (Hz), L is the length of the string (m), T is the tension in the string (N) and μ is the linear density (mass per unit length, kg m^{-1}).

Assuming that the G string of a violin has a length of 60.0 cm, the tone of G corresponds to a resonant frequency of 196 Hz, and that the mass of the string is approximately 0.900 g, we can use Mersenne's equation to show that the tension in the string is approximately 83.0 N. This corresponds to the string having about 8.50 kg hanging off it.

Physics file

The combined tension in the strings of a full-sized grand piano is equivalent to 20 tonnes sitting on top of it!

Figure sm.7 Didgeridoo player.

See Resonance on page 204.
See Harmonics on page 207.
See Wind instruments and air columns on page 209.

Brass and the didgeridoo

Didgeridoo players (and trumpet players) will tell you they generate the sound by vibrating their lips within the mouthpiece of their instrument. The air within the tube acts as the resonator, and it is the length of the tube that generally dictates the pitch of the instrument. It is the opening at the end of the didgeridoo that acts as the radiator. Typical brass instruments, such as trumpets, have a bell at the end which radiates the sound more efficiently.

Maximum variation in pressure produced by the player's vibrating lips (and a corresponding minimum displacement of the air particles there) occurs at the mouthpiece of the didgeridoo. This corresponds to a pressure antinode or a displacement node.

The open end of the didgeridoo corresponds to a pressure node (and a displacement antinode). Hence the didgeridoo and brass instruments, such as the trumpet and trombone, are closed air columns and the air within them can vibrate at their fundamental frequency and their overtones are odd harmonics. So if f_1 is the fundamental frequency then the first overtone (third harmonic) is f_3 and the second overtone (fifth harmonic) if f_5 and so on. The frequency of the third harmonic is three times the frequency of the fundamental frequency, while the fifth harmonic is five times the frequency of the fundamental. The fundamental mode of vibration occurs when the effective length of the closed air column is a quarter of the wavelength of the fundamental frequency. The waves generated by conical brass instruments, such as the French horn and the tuba, are somewhat more complicated. The players of these instruments are able to generate both odd and even harmonics by the appropriate placement of their hands in the bells of their instruments.

✓ Worked Example sm.3

It is a hot day of 36.0°C and a busker is playing his didgeridoo outside the Fremantle markets. If the didgeridoo is 120 cm long, what will be the frequency of its fifth harmonic?

Solution

The didgeridoo is a closed air pipe; hence the wavelength is 4 times the length of the didgeridoo:

$L_{didg} = 120 \times 10^{-2}$ m

$$\lambda = 4L = 4(120 \times 10^{-2}) = 4.80 \text{ m}$$

Since the temperature is 36.0°C, the speed of sound:

$T_{air} = 36.0°C$

$$v_{sound} = 331 + 0.6T = 331 + 0.6(36.0) = 3.53 \times 10^2 \text{ m s}^{-1}$$

Hence, the fundamental frequency of vibration for the didgeridoo is:

$v_{sound} = 3.53 \times 10^2$ m s^{-1}

$\lambda = 4L = 4.80$ m

$$f_1 = \frac{v}{4L} = \frac{3.53 \times 10^2}{4.80} = 73.5 \text{ Hz}$$

Hence, the fundamental frequency of vibration for the didgeridoo is 73.5 Hz. Notice how low this is. A closed pipe will only produce odd harmonics:

$f_1 = 73.5$ Hz

$$f_5 = 5 \times f_1 = 5 \times 73.5 = 367 \text{ Hz}$$

Woodwinds

The frequencies generated by brass instruments are principally determined by the player's vibrating lips. With woodwind instruments, the player either blows at an edge, as with a flute, or into a reed (either single, as with a clarinet, or double, as with an oboe).

Edge woodwind instruments, such as recorders and flutes, generate vibrations by making the air around the edge swirl around like a flag does in the wind. This effect is called turbulence. This turbulence is rhythmic enough to make the air resonate in the instrument. Reed instruments work in a similar way but this time the flag is the reed. Reed instruments do not waste or lose air, so they tend to be louder than edge instruments.

The holes along the length of wind instruments effectively allow the player to change the length of the air column and hence the frequencies at which the instrument resonates. The sound radiates out of the end of the instrument and, to a smaller extent, out of the open holes.

Woodwind instruments and reed instruments like the saxophone and oboe are considered to be like a tube open at both ends. The fundamental mode of vibration of this type of air column occurs when the effective length of the open pipe is half of the wavelength of the fundamental frequency. For these types of instruments the overtones are both odd and even, so if f_1 is the fundamental frequency then the first overtone (second harmonic) is f_2, the second overtone (third harmonic) is f_3 and so on.

Figure sm.8 Flautist and clarinetists performing together.

Harmonic spectra

When a wave is generated it isn't only the fundamental mode of vibration that occurs in an instrument. Depending on the instrument, and the skill of the player, a range of harmonics will be present. This characteristic mix of harmonics (or overtones) is called the instrument's *harmonic spectrum* and gives rise to the instrument's characteristic sound or timbre.

It isn't easy to imagine a string, for example, oscillating in its fundamental mode and its second harmonic simultaneously, but this is what happens. The air within wind instruments (brass and woodwinds) also vibrates as a mixture of harmonics. Physicists can at least picture this using the law of superposition. Figure sm.10 compares the complex waves produced by a violin, piano and recorder. Notice that not all the frequencies or harmonics are equally strong or loud (i.e. they have different amplitudes).

This brings us to electronic keyboards. The complex wave patterns of a large range of instruments have been analysed to allow electrical engineers to design keyboards that can mimic an instrument of choice. To achieve this they have used a process called Fourier analysis. This is a process whereby a complex waveform is decomposed into a series of basic sine wave forms with varying amplitudes and frequencies. It is like using the law of superposition in reverse.

As an example, Figure sm.9 illustrates, in more detail, the harmonics that are present when a flute plays a concert A, the tuning note for an orchestra, which has a standard frequency of 440 Hz. When an electronic keyboard is made to play like a flute it is programmed to choose the appropriate frequencies and amplitudes to produce the required sound.

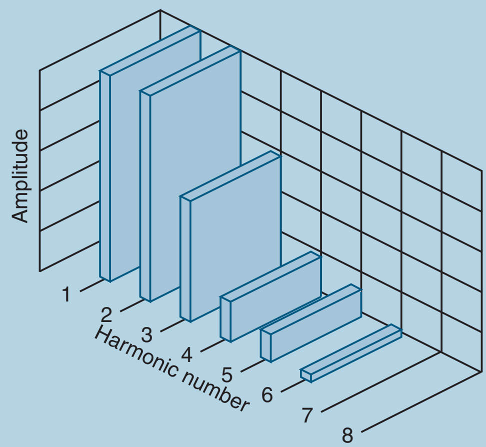


Figure sm.9 The harmonic spectrum of a flute sounding a concert A.

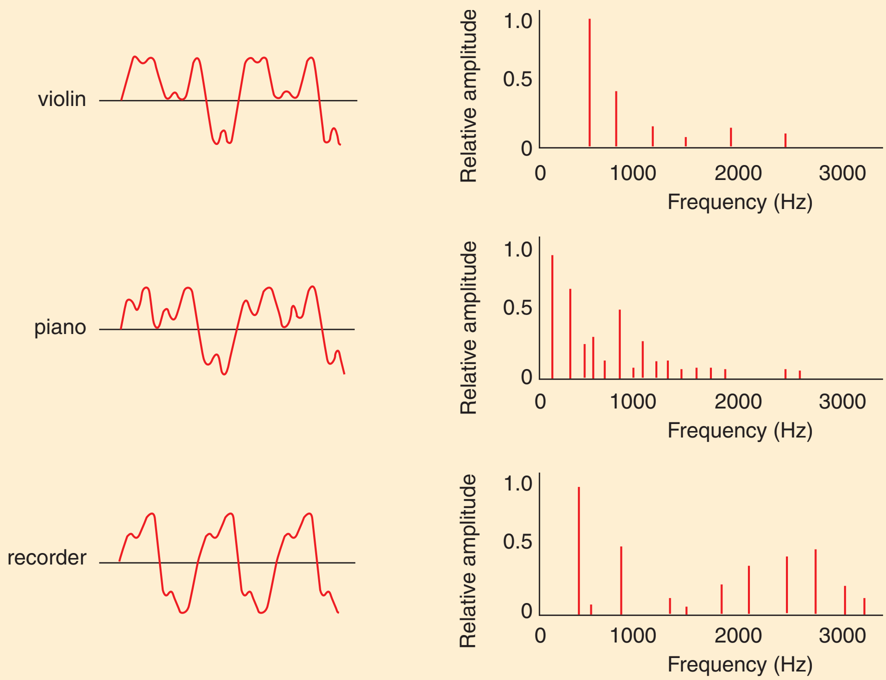


Figure sm.10 The complex waveforms of the violin, piano and recorder result from the superposition of a range of frequencies with differing energies and duration.

Investigation

1. Discuss what the generator, resonator and radiator would be for a trumpet, a clarinet and a piano.
2. Why does a bass guitar create lower notes than a violin (the answer is not only its length)?
3. How can you increase the pitch a string produces if you cannot increase the tension in the string?

The voice

There are good reasons why computer-generated voices sound very unnatural. Biology is far more complicated and subtle than physics. Though many of us are timid when it comes to singing, whatever sound we produce when we do sing, the associated waveform is very complicated and difficult to analyse.

The remarkable thing about the voice is that it can produce such a vast range of frequencies (or tones) with vocal cords (the vibrating strings) that are relatively short. The vocal tract (i.e. the length from the voice box, or larynx, to the mouth) is just as important in producing the sound. The vocal tract, in effect, acts as a radiator, like the belly of a violin.

Exercises

E3 Use Mersenne's law of stretched strings to determine the factor by which the frequency of a string will change when the tension in it is tripled.

E4 What would be the fundamental frequency of vibration of a 40.0 cm string that has a tension in it of 120.0 N and a mass of 30.0 g?

Exercises

E5 A student wants to analyse the sound made from an open and a closed pipe by comparing the sound spectrum of the same 100.0 Hz note played by each instrument.

- **a** What is the shortest length of open pipe needed in order to produce a note of frequency 100.0 Hz at a temperature of 26.0°C?
- **b** What shortest length would be needed if a closed pipe is used?

E6 An experimenter blows across the top of a tube, creating standing waves within the tube. Since the lips are not covering the tube this is an open tube. Calculate the frequency of the third harmonic if the tube is 40.0 cm long and the speed of sound is 340.0 m s^{-1}.

Acoustics

Those of us who leave our singing only for the shower know that it sounds rather loud. We are aware that it is the hard surfaces in the bathroom that cause this. The physics term for what is occurring is **reverberation**. The sound 'appears' loud simply because it takes longer to die away as there are few absorbing surfaces.

Those who design concert halls practise an art. This is to say that while there is no 'exact' science of acoustics, designers do have to take into consideration many physics phenomena.

Reflections

When candles were the sole source of light at night, mirrors were used to double the available light. The same principle is used at outdoor auditoriums where there is no electronic amplification. Band shells are designed to project all of the available sound towards the audience.

One problem with this approach, however, is that there is distortion because the reflecting shell is more effective for higher frequencies than for lower frequencies. This is due to the phenomenon of diffraction. If the wavelength of a sound is greater than the size of the object it is supposed to reflect off, it will tend to go around it instead.

See Reflection of waves on page 192.

Practical activity

21 The speed of sound by clap and echo

Figure sm.11 Band shells are used in outdoor auditoriums to reflect the sound waves towards the audience.

✓ Worked Example sm.4

A partition 4.00 m wide is placed behind a band. Estimate the range of frequencies that will be effectively reflected from its surface.

Solution

Reflection of sound will occur for wavelengths equal to or less than 4.00 m. Assume the speed of sound is 346 m s^{-1}:

$v_{air} = 346 \text{ m s}^{-1}$

$\lambda = 4.00 \text{ m}$

$$f = \frac{v}{\lambda} = \frac{346}{4.00}$$
$$= 86.5 \text{ Hz}$$
$$= 90 \text{ Hz}$$

Hence, the corresponding frequency is about 90 Hz. Since smaller wavelengths correspond to higher frequencies, it follows that the partition could reflect all frequencies above about 90 Hz. Note that answers to estimation questions should only be expressed with one significant figure.

The answer to Worked Example sm.4 may seem adequate, but this means that the bass notes will undergo less reflection and the audience will hear a distorted sound because the higher frequencies will be more prominent.

Physics file

Bats are well known for using reflection or echoes to find their way around in the dark. This is termed echolocation. Bats simply project very high frequency sound pulses and 'time' how long it takes for the echo to return. As they get closer they give out more pulses per second to increase the detail of their 'sound picture'. This is how they judge their distance from obstacles and potential food!

Figure sm.12 Bats have large ears that help them 'collect' the echo of the pulse they emit.

Investigation

Find out why bats use such high-frequency sound to echolocate. Look at the question quantitatively as well as descriptively.

Interference

In enclosed halls the sound can not get out—it is like your bathroom, only bigger. Too much reverberation can cause problems. The problem with these multiple reflections is that the sound hangs around and interferes with the latest sound being projected from the performers. This is why sports halls, for example, do not make very good music-performance venues: the music sounds 'noisy'.

Consider the following observation. A pianist was performing a Beethoven piano concerto. The pianist had a bad habit of keeping his own beat by tapping the floor with the heel of his shoe (he did not seem very interested in the conductor). A physics teacher is a member of the audience and is seated near the front. He observed that the noise from the heel seemed rather loud. He came to the conclusion that he must have been sitting at a point of constructive interference. He heard the heel 'directly', but also from reflection off a wall at the side of the stage. Worked Example sm.5 illustrates this point.

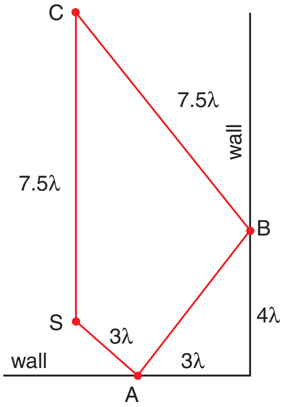
Figure sm.13 The inside of a concert hall is designed to have excellent acoustics.

✓ Worked Example sm.5

Sound waves of wavelength λ pass from a source S (see the figure below). The sound reaches an observer at C by two paths: directly and by reflection from the two walls. All of the distances are indicated on the diagram as multiples of the wavelength of the sound.

a What is the path difference of the waves meeting at C?

b Would the observer hear a maximum or minimum sound? Give a reason for your answer.



Solution

a Using Pythagoras's theorem:

$$AB^2 = (3\lambda)^2 + (4\lambda)^2$$

$$AB = \sqrt{9\lambda^2 + 16\lambda^2} = \sqrt{25\lambda^2} = 5\lambda$$

$$Path_1 = 7.5\lambda$$

$$Path_2 = 3\lambda + 5\lambda + 7.5\lambda = 15.5\lambda$$

$$\text{Path difference} = 15.5\lambda - 7.5\lambda = 8\lambda$$

b Since this path difference is a whole multiple of wavelengths there will be constructive interference between the sound waves at C, so the sound will be louder than normal.

See Superposition on page 202.

Practical activities

17 Interference of sound
18 Interference of water waves

Extended response

Would there be other places where the sound from the pianist's shoe would be 'loud'? Give a clearly justified 'physics' answer.

Activity

Investigation

Purpose

To observe and record the effects of interference of sound.

Materials

- two speakers
- a signal generator
- a sound level meter
- 30 m measuring tape
- power source

Procedure

Set up the speakers in a large room (preferably with curtains around it) or preferably in a hall with the speakers not close to any walls, and investigate the interference of sound. First walk with the sound level meter across the path of the speakers and observe the variation in intensity.

Discussion

In what ways do your observations change as you vary the orientation and/or distance between the speakers? Compare the interference patterns formed by different frequency sounds.

Practical activity

22 Ultrasound interactions: Attenuation of sound

Absorption

The degree to which a surface will absorb sound is dependent on its nature (its size and the material from which it is made), and also on the wavelength of the incident sound. This is, again, related to diffraction. The effective absorbing area of a surface can be calculated by multiplying the true area by an absorption coefficient. Table sm.2 gives the coefficients for a variety of surfaces and frequencies of sound.

table sm.2 Absorption coefficients of standard materials for sounds of different frequencies

Material	Frequency (Hz)				
	125	250	500	1000	2000
Marble or glazed tile	0.01	0.01	0.01	0.01	0.02
Brick wall	0.01	0.01	0.02	0.02	0.02
Wooden floor	0.15	0.11	0.10	0.07	0.06
Plywood on struts	0.60	0.30	0.10	0.09	0.09
Plastered walls	0.30	0.15	0.10	0.05	0.04
Carpet with underlay	0.08	0.27	0.39	0.34	0.48
Heavy curtains	0.14	0.35	0.55	0.72	0.70
Cane tiles on concrete	0.22	0.47	0.70	0.77	0.70

Higher coefficients imply that greater absorption occurs at that frequency. Notice that soft surfaces (carpets, curtains) absorb higher frequencies very well. This is because the corresponding shorter wavelengths get caught. Why would a wooden floor be a good absorber for lower frequencies?

These numbers are used to determine an approximation for the reverberation time defined as the time it takes for a sound to die away.

Knowing the reverberation time for a specific hall is important for deciding what type of music is suitable for performance within it.

Finally, in modern concert halls consideration is made to differences in reverberation time caused by the size of the audience. To maintain a relatively stable reverberation time within a hall, the chairs are made and lined with soft materials that mimic the absorbing properties of clothed bodies. If the hall is half empty the music will then sound roughly the same as if it were full. The chairs are not designed just to be comfortable!

Investigation

1. Compare the reverberation times for your science laboratory with those for a room with curtains and carpet.
2. Soundproofing can be for the purpose of shielding against unwanted sound (such as the city sounds outside of a concert hall) or protection from harsh and damaging sound (such as from the crack of a starter pistol). Investigate the soundproofing techniques employed in these two instances.

Exercises

E7 Estimate the size of a band shell in order for it to reflect all frequencies above 60.0 Hz.

E8 A bat is hunting in its cave where the temperature is 2.00°C. The bat lets out a pulse of sound of frequency 150 000 Hz. It takes 0.0300 s for the pulse to return. How far away is its prey?

Effects and uses of sound

Music is important to most of us, even if your music is not appreciated by everyone at home. There are, however, a variety of practical uses for sound. To understand how these applications work, we first need to look at the idea of energy with respect to sound.

You may be familiar with a device known as a 'Newton's cradle'. It comprises a row of five balls hung using fishing line. The two end balls can be made to bounce backwards and forwards while the centre balls remain (almost) stationary.

Figure sm.14 Newton's cradle is an 'executive toy' based on the principle of conservation of momentum extended over a number of objects.

You may be aware that the stopping of one ball and the moving-on of the next is possible because the balls are all the same size. This enables the kinetic energy of each ball to be passed on to the next without any 'reflection' of energy. The same effect can be observed on a pool table. If, however, the masses were different, then the first ball would have a final velocity other than zero. Let us assume we have a 'two ball' Newton's cradle. By using the laws of conservation of momentum and kinetic energy, it can be shown that the percentage of energy reflected when the first ball hits the second stationary ball is given by:

$$E_R = 100\left(\frac{m_1 - m_2}{m_1 + m_2}\right)^2$$

where E_R is the reflected energy (%), m_1 is the mass of the first ball (kg) and m_2 is the mass of the second stationary ball (kg).

See Conservation of energy on page 435 on CD.

The equation above gives the ratio of the first ball's final kinetic energy to its initial kinetic energy as a percentage. Note that, as observed, if $m_1 = m_2$, then this ratio is zero, and hence all the energy from the first ball is transmitted, or passed on, to the second ball.

✓ Worked Example sm.6

A ball rolls into a stationary ball of three times its mass. What percentage of its initial energy will it have after the collision?

Solution

$m_1 = m_1$

$m_2 = 3m_1$

$$E_R = 100\left(\frac{m_1 - m_2}{m_1 + m_2}\right)^2$$
$$= 100\left(\frac{m_1 - 3m_1}{m_1 + 3m_1}\right)^2$$
$$= 100\left(\frac{-2m_1}{4m_1}\right)^2$$
$$= 100(-0.5)^2$$
$$= 25\%$$

Hence, the ball will have 25% of its initial energy after the collision.

Extended response

Going further

Using the laws of conservation of momentum and kinetic energy, show that the percentage of energy reflected when a ball of mass m_1 hits a second stationary ball of mass m_2 is given by the equation shown above. Assume that the initial speed of the incident mass, m_1, is a and its final speed is b, and that the speed of m_2, initially at rest, is c after colliding with m_1.

Impedance

The term *impedance* in physics comes from the word 'impede', which means to hold back. Acoustic impedance refers to resistance to the transmission of sound through a medium. It is analogous to the resistance to the flow of electrons in a wire and is also measured in ohms. When sound meets a boundary between two media (e.g. air and water), there is an impedance mismatch and not all of the signal will be transmitted. It so happens that the percentage of energy reflected at such a boundary is given by:

$$E_R = 100\left(\frac{Z_1 - Z_2}{Z_1 + Z_2}\right)^2$$

where E_R is the reflected energy (%) and Z_1 and Z_2 are the impedances of the media the wave is passing from and to, respectively (Ω).

The similarity of the last two equations should not be a surprise as mass impedes motion due to its inertia. Acoustic impedance is important in the operation of wind instruments and in the application of sound to ultrasound imaging and mining.

How is it that standing waves can be set up in a tube (e.g. a clarinet) given that they rely on a reflection of the wave at the open end of the instrument? The answer is impedance. The impedance of air to longitudinal waves passing through it is dependent on the cross-sectional area. Since the cross-sectional area of the clarinet is different from that of the outside air, there is an impedance mismatch. Therefore, some of the energy of the wave is reflected, allowing for the formation of the standing wave within the instrument. Clearly some of the sound does get out; otherwise we would not hear the instrument!

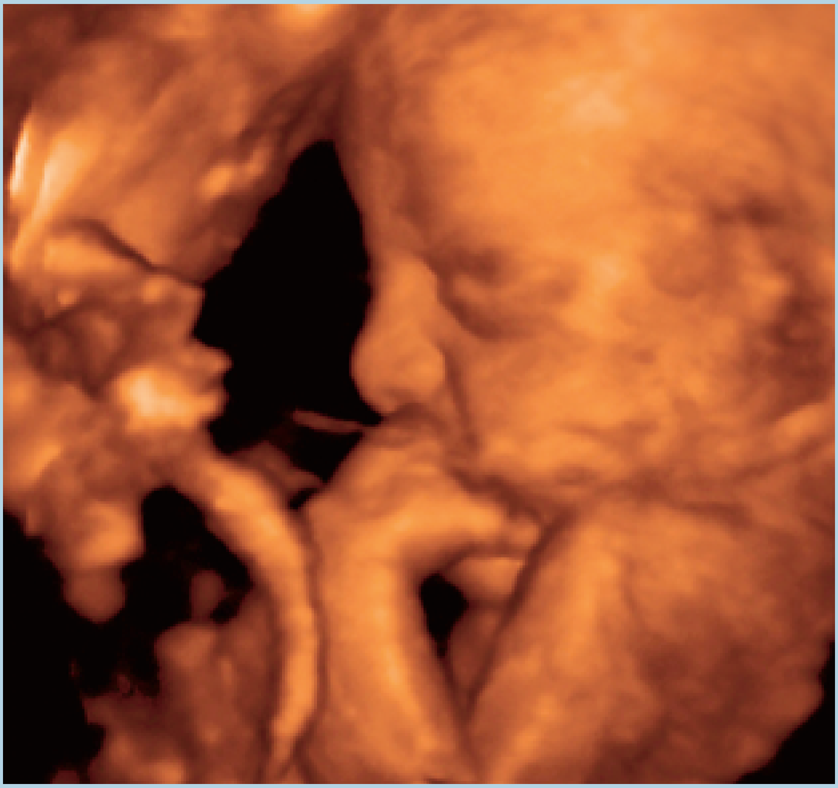


Figure sm.15 Ultrasound image of a baby in utero.

Physics file

You may have seen the images created of unborn babies while they are still in their mother's uterus. These images are created by directing very high frequency sounds through the wall of the mother's uterus. Every time the sound meets a new type of tissue (with its corresponding acoustic impedance), there will be an impedance mismatch and some of the signal will be reflected while the rest will be transmitted. The machines attached to the probe pick up the reflections or echoes. This information is used to generate the image we see (Figure sm.15).

Investigation

Seismology is the use of sound to investigate what is under the ground. Research how acoustic impedance is important to the mining industry.

Tuning

When musicians prepare for performance, the most important thing they must do is ensure that they are 'in tune', at least with one another. Several previous examples have illustrated how the speed of sound in air is dependent on the air's temperature. This is particularly important for wind players as the frequency that their instruments produce depends strongly on the length of the air column in their instrument, and hence the wavelengths of waves that can be generated within them.

✓ Worked Example sm.7

Minh warms up his flute ready for an Anzac Day performance with his band, which is to be held outside. It was 25°C in the room where he warmed up, but it is 40°C outside. How out of tune will his concert A (440 Hz) be, and what can Minh do about it?

Solution

Since the length of Minh's flute has not changed, the wavelength at which the air within it will resonate has not changed. Because the temperature of the air has changed, the frequency at which these resonant wavelengths will occur will be slightly different.
Since λ is constant, it follows that:

$v_{25} = 346 \text{ m s}^{-1}$

$v_{40} = 331 + 0.6(40) = 355 \text{ m s}^{-1}$

$f_1 = 440 \text{ Hz}$

$$\lambda_1 = \lambda_2$$

$$\frac{v_1}{f_1} = \frac{v_2}{f_2}$$

$$f_2 = \frac{v_2 f_1}{v_1} = \frac{(355)(440)}{346}$$

$$= 451 \text{ Hz}$$

Hence, Minh's concert A now has a frequency of 451 Hz, which will be noticeably sharp. What all wind players do in this situation is physically change the length of their instrument's air column. For Minh to lower the pitch of his instrument, he will have to increase the length of the flute by pulling the head of the flute out. Increasing the length of the air column will increase the wavelength and hence decrease the frequency at which it will resonate.

So how does a group of musicians get in tune with one another? They achieve this by listening for and then eliminating beats by adjusting their instruments appropriately. Beats are heard whenever two sounds that are close in frequency are played together. What is heard is an oscillation in the volume of sound. Figure sm.16 illustrates the waveform that is formed by two sounds of close frequency.

When the frequencies are very close the oscillation will be slow. As the frequencies become more separated, the oscillation increases until eventually the listener hears two distinct tones.

The rate of the oscillation is called the beat frequency, and is simply the difference between the two separate frequencies, as defined by the equation below. Musicians know they are in tune with one another when beats are not heard. Note that a practised musician is able to determine whose instrument is sharp (above the required pitch) and whose is flat (below the required pitch), and they adjust their instrument accordingly, hence eliminating the beats.

$$f_{beat} = |f_1 - f_2|$$

where f_{beat} is the beat frequency (Hz) and f_1 and f_2 are the two separate frequencies (Hz).

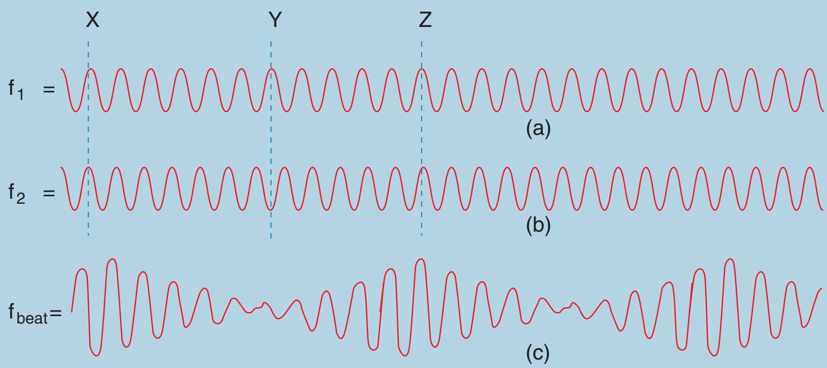


Figure sm.16 A beat (f_{beat}) produced by two sound waves, f_1 and f_2, of equal amplitude but slightly different frequency. At time X the waves are in phase, producing a larger amplitude. A short time later, at Y, they are out of phase, resulting in a smaller amplitude, and at Z they are in phase once again. When the two waves are in phase (at X and Z), the resulting pressure variation, and hence volume, is large. At Y the waves are out of phase and the resulting amplitude and volume are zero.

Experimental investigation

Research methods used to measure acoustic impedance and investigate the acoustic impedance of a variety of materials, such as polystyrene, sheets of plywood, curtain material. How, if at all, is the acoustic impedance of a material dependent on the frequency of the incident sound?

Exercises

E9 Why do wind and brass players keep their instruments warm by blowing through them?

E10 A violinist deems her violin to be sharp (higher than it should be). What will she do to the string to get it in tune?

E11 A ball hits a stationary ball of half its mass. What percentage of the incident ball's energy will the second ball have after the collision?

E12 The formula for calculating the percentage of energy reflected is also applicable to light, as defined by:

$$E_R = 100\left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2$$

where E_R is the reflected energy (%) and n_1 and n_2 are the refractive indices of the media the light is passing from and to, respectively.

Use this to determine what percentage of light is reflected when it passes from air ($n = 1.00$) into water ($n = 1.33$).

E13 Yuki is tuning her violin by comparing the note she is playing with a tuning fork of 294 Hz. She hears about 3 beats per second. Explain what Yuki should do to tune her violin string.

Loudspeakers

For most of us, the music we hear comes from speakers. There are MP3 player headphones, CD player speakers, bookshelf speakers for mini-hi-fi systems, and the large speakers that stand on the floor, on which serious audiophiles can spend lots of money. There are also the very large speakers that rock bands use.

The design of these speakers is dependent on the means by which electrical signals are used to generate the movement of the speaker cones, but there are many properties of sound waves that influence the size and physical structure of the speakers. This final section will look at what dictates the design of speakers, and we will see how any particular speaker is a compromise of engineering and cost constraints.

Why so many drivers?

It is impossible for a single speaker (except for expensive headphone speakers) to produce quality sound at loud listening levels across the whole audible spectrum. A serious hi-fi loudspeaker is a box (the enclosure) that houses, generally, three speakers (or drivers). Many of you will be aware that these drivers are called the woofer, the mid-range and the tweeter. Three drivers is the most efficient number of drivers for meeting our ear's sensitivity to different frequencies of sound.

Woofers are generally large for two reasons. First, much of the acoustic, or sound, energy in music is found in the lower range of frequencies (say, below 500 Hz). Second, looking at the graph in Figure sm.4 reminds us that our ears are less sensitive to this range of frequencies. Therefore the woofer needs to be physically bigger to push a greater quantity of air, hence producing a more intense sound.

If the mid-range driver was the same size as the woofer, then the listener would experience distortion; that is, the mid-range frequencies (~500→5000 Hz) would be too prominent. Mid-range drivers can be smaller because our ears are more sensitive to these frequencies. A similar explanation applies to the generally tiny tweeter. The tweeter can also be small because not much music uses this frequency range (>5000 Hz).

Figure sm.17 A hi-fi speaker usually has three drivers. From top to bottom they are called: the tweeter, the mid-range and the woofer.

This is fortunate because the tweeter has to oscillate backwards and forwards up to 15 000 times per second! Clearly, to achieve this, an object with little inertia (i.e. mass) is preferable.

The woofer

The job of the woofer very much dictates the size and general construction of a loudspeaker. Here we will look more closely at acoustic demands and limitations and how they are reconciled.

The vast majority of drivers are permanent magnet speaker cones. Figure sm.18 illustrates their basic construction. The proportions will be different for the various sizes of driver, but they are similar in construction.

The essential components of the driver perform the following functions:

- The diaphragm or cone is responsible for pushing the air.
- The flexible edge or suspension holds the diaphragm in place.
- The voice coil has the electrical wire from the frequency crossover network wound around it.
- The spider holds the voice coil centrally.
- The permanent magnet interacts with the voice coil, causing the diaphragm to oscillate.

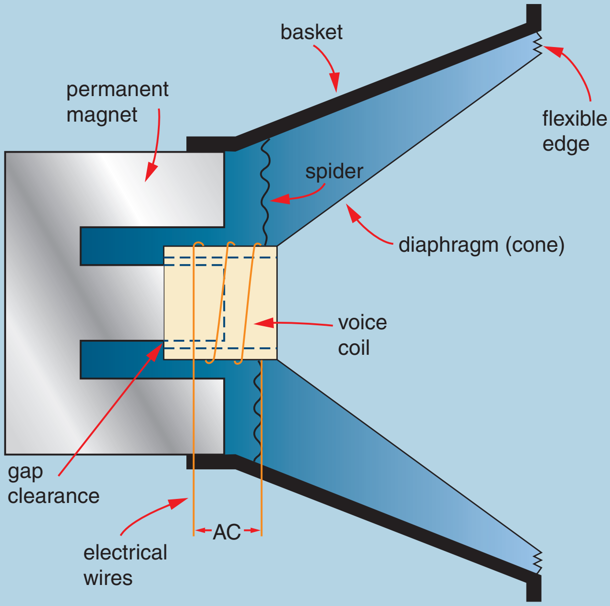


Figure sm.18 Cross-section of a woofer driver. The components of other drivers are similar.

The resonant frequency of a woofer, as determined by the mass of the diaphragm and the elasticity of the suspension, imposes a lower limit on the frequency of sound the woofer can effectively reproduce. Ideally this would be set as 20 Hz, but this is not practical. A typical woofer may have a resonant frequency of 30 Hz, meaning the driver will effectively produce sound above 35–40 Hz. If the woofer does try to produce sound around 30 Hz there will be a resonant peak and the sound will boom. Some listeners might like this effect, but it isn't in the best interests of the life span of the woofer!

The permanent magnet of the driver provides a magnetic field that interacts with the continually changing magnetic field created within the voice coil. The voice coil's magnetic field is a consequence of the variable AC signal that flows though the wires that are wound around it. The whole process is a consequence of magnetic forces.

See 3.2 Force on current-carrying conductors on page 104.

The efficiency with which the electrical energy of the AC signal is converted to the vibrational energy of the diaphragm is dependent on minimising the size of the gap between the voice coil cylinder and the inner cylinder of the permanent magnet. The gap can range from 8 to 25 thousandths of a centimetre. The spider is paramount in providing the necessary stability and centrality of the voice coil.

The enclosure

There are specific physical reasons as to why the drivers are enclosed within a box; it is not just for convenience or so they look good. Again, it is the woofer that most needs the enclosure and there are various reasons for this, relating to interference and diffraction.

Woofers produce the lowest notes. We know that lower frequency sound waves are more prone to diffraction than higher frequency waves; that is, they bend around corners better. The woofer must be held within an enclosure to reduce destructive interference. Without the enclosure the cone moves forward, producing a compression; part of this wave would spread around behind the woofer and destructively interfere with the rarefaction that has been produced at the back of the cone. The net effect would be a decrease in the intensity of the emitted sound. The cabinet stops this from occurring and acts as a baffle. Note that sound is emitted from the rear surface of the diaphragm, not just the front.

The size of an enclosure is dependent on the size of the woofer it must house. Sealed enclosures accept that the rear directional sound wave from the diaphragm will be 'lost'. Soft material is placed within the enclosure to absorb its energy rather than have it reflect off the back wall of the enclosure and interfere with subsequent sound waves. The problem with sealed enclosures is that the air trapped within the enclosure acts as an acoustic spring against the movement of the diaphragm. It indirectly decreases the elasticity of the suspension, thus increasing the woofer's resonant frequency, which subsequently decreases its ability to produce the very low notes. To compensate for this, sealed enclosures tend to be quite large as this reduces the springiness of the enclosed air.

There is a way of reducing the size of the enclosure. Since the entrapped air acts as a spring, we can use it to good effect and increase the elasticity of the suspension, i.e. make it softer. This allows us to decrease the size of the diaphragm (its mass) and, hence, the size of the driver. This, however, poses a problem. If the diaphragm is smaller in diameter, it won't be able to push out as much air and the sound intensity will be reduced. To compensate for this, the drivers are designed to have a long throw; that is, the speaker cone oscillates with greater amplitude, hence the term 'long throw'. Such loudspeakers are called acoustic suspension loudspeakers and they are able to be physically smaller, hence the bookshelf speaker.

Rather than waste it, vented enclosures attempt to make use of the rear directional sound wave that is emitted from the back surface of the diaphragm. A vent or port lets the sound out the front or, in some cases, the back of the enclosure. But here there is another potential problem. The baffle was there to prevent destructive interference from occurring but now that possibility re-emerges. To prevent this from occurring, the enclosure designers choose dimensions that cause the rear directional sound wave to emerge from the vent after it has traversed half a wavelength. Note that it must be a half wavelength since the forward and rear directional waves are 90° out of phase when first produced. This means that the two waves will constructively interfere and the intensity level increase. This also enables the size of the woofer to be reduced.

Horn enclosures are similar in effect to the bell of brass instruments. Rather than have a large diaphragm, the driver is placed in a small enclosure and a horn is attached that projects the sound out into the room more effectively owing to the gradual change in acoustic impedance from the diaphragm to the air in the room.

The sound from such horns is able to spread out or disperse more effectively. Woofers do not need horns as their emissions have low frequencies that diffract easily. This is just as well because an effective horn for a typical woofer would have to be 3–4 m in diameter at its mouth. Tweeters also use dome-shaped diaphragms to increase the dispersion of their high-frequency emissions.

*
See 5.1 Wave properties on page 184.
See Diffraction of sound on page 197.
See 5.3 Wave interactions on page 201.

Specifications

When you are looking to purchase a speaker, there are various specifications which can help you decide. In the end, however, it will be your ear (i.e. what you hear) that will dictate your final choice. The important specifications are:

- Sensitivity gives the sound intensity level at a certain distance from the speaker with a specified power input; for example, 85 dB @ 1 m with 1 W input.
- Frequency response specifies the variation in sound intensity across the frequency range of the speaker. It is sometimes illustrated using a frequency response curve; for example, 40 Hz → 16 000 Hz, ±4 dB.
- Speaker dispersion is a measure of the diffraction properties of the speaker. It specifies the angle through which the sound is effectively dispersed and the variation in intensity; for example, 120° dispersion, ±6 dB from 50 Hz → 16 000 Hz.
- Speaker impedance is the effective resistance of the loudspeaker to the AC signal from the amplifier. The most common impedance is 8 Ω. Some speakers have a choice of rear sockets that allows them to be run at 4 Ω. In this case the speaker will draw more current from the amplifier, and will tend to be more responsive to peaks in sound level.
- The final important loudspeaker specification is its power rating. There may be a minimum and maximum power rating given. If the minimum power rating is 30 W, the amplifier must provide at least this amount per channel (or speaker). If less than this amount is provided, 'clipping' can occur where high harmonics, which cause damage to the tweeters, are generated. Efficient speakers are those that have a low minimum power rating.

Some of us enjoy loud music, but to achieve this a powerful amplifier is required. Clearly though, we must be careful not to burn out the loudspeakers. Most speakers will specify a maximum power rating. If the maximum rating is 60 W, it should not have an input of more than 55 W. Note that the maximum power rating is a measure of the power consumed, not the acoustic energy produced. Loudspeakers are rarely more than 10% efficient. The efficiency of a loudspeaker is defined by:

$$\%\text{ efficiency of a loudspeaker} = \frac{\text{output acoustic power} \times 100}{\text{input electrical power}}$$

Figure sm.19 An old-style gramophone used a horn to project the sound around the room.

Exercises

E14 Which of the following speakers is most efficient?

	Amp power to speaker (W)	Sound power from speaker (W)
A	200	2.00
B	100	10.00
C	20	1.60
D	5	0.20
E	2	0.02

E15 You own an amplifier with a 40 W per channel power rating. Which of the following speakers should you buy?

	Minimum recommended power rating (W)	Maximum recommended power rating (W)
A	20	80
B	50	120
C	10	35
D	35	90
E	5	30

E16 An 80.0 W per channel amplifier is driving a pair of 8.00 Ω speakers. Determine the maximum current through the connecting wires.

Investigation

1 Investigate how small diameter woofers generate the same loudness level as larger woofers.
2 Find out about the generation of sound produced by a bass reflex loudspeaker.
3 Why are loudspeakers rarely more than 10% efficient?
4 Research how the combination of inductors and capacitors is used in the frequency crossover network.

5 Wave properties and light

Sound permeates the physical world. It adds an extra dimension to our experience of the world around us, from the rustle of leaves in the wind to the sheer noise and vibration of a jet aircraft thundering off the runway. Sound is important to many forms of communication for humans and other animals. For many, it is the major way of receiving information about the world around them. As music, it is a form of entertainment, lifting the spirit and allowing a depth of expression rivalled in few other fields.

This study introduces some simple acoustic concepts that govern how we hear sound, along with the theory behind the production of real, everyday sounds through natural and manufactured means. Some of the early work may be familiar to you from earlier studies of waves. Though the concepts covered are relatively simple, a sophisticated understanding of acoustics is required to reproduce sounds clearly.

All of our great physicists had something of the pioneer and the radical in them. Each came across observational data that either defied established theory or for which there was no existing explanation. They were willing to think outside the established approaches of their time, break new ground and develop their own new theories.

The development of scientific knowledge has run parallel with technological development. As our ability to produce more sophisticated measuring equipment grows, new areas of investigation are opened up to us. We have even become knowledgeable enough to realise that by our acts of observation themselves we ineradicably influence the things we see. As our understanding of light develops in the future, we will continue to need pioneering radical physicists to create new models that will help us to understand the fascinating phenomenon of light.

By the end of this chapter

you will have covered material from the study of wave properties and light including:

- the longitudinal wave nature of sound (pitch, frequency and period, amplitude, wavelength, speed)
- the diffraction of sound through gaps and around obstacles
- the superposition of waves from more than one source
- standing waves and resonance
- the photo-electric effect as evidence for the particle-like nature of light and counterevidence for the wave model of light.
- the de Broglie wavelength
- atomic absorption and emission spectra including vapour lamps
- changes in energy levels of atoms and photon energy, frequency and wavelength
- the quantised atom and standing waves as a model of the dual nature of matter.

5.1 Wave properties

Waves

Most of us will be familiar with wave motion of one type or another. While the existence of ocean waves and the ripples in a pond are accessible visible examples, you may also have some background knowledge of sound waves, waves in stringed instruments, radio waves or microwaves. All waves involve **energy** being transferred from one location to another, without any net transfer of matter. Using the easily visualised waves in water and springs, we shall first summarise the properties common to all kinds of wave phenomena. Later you will rely on this knowledge to further develop your understanding of the nature of light.

Waves can be grouped into two major categories: those which rely upon an elastic medium to carry them, or **mechanical waves**, and those which do not require a medium. We will begin with an examination of mechanical waves in springs.

Consider a slinky spring that has been gently stretched, laid on a smooth surface and held at each end by a student. If one student were to give a quick sideways movement to one end of the spring, a single **wave pulse** would be seen travelling the entire length of the spring (Figure 5.1). Hence, energy would travel away from the source of the disturbance. The item carrying the disturbance, in this case the spring, is called the **medium**. The particles of the medium each undergo a vibration as the wave energy passes through. When the displacement of the particles is at a right angle to the direction of travel of the wave, the wave is called a **transverse wave**. The **amplitude** of the wave is defined as the maximum displacement that a particle has from its original rest position.

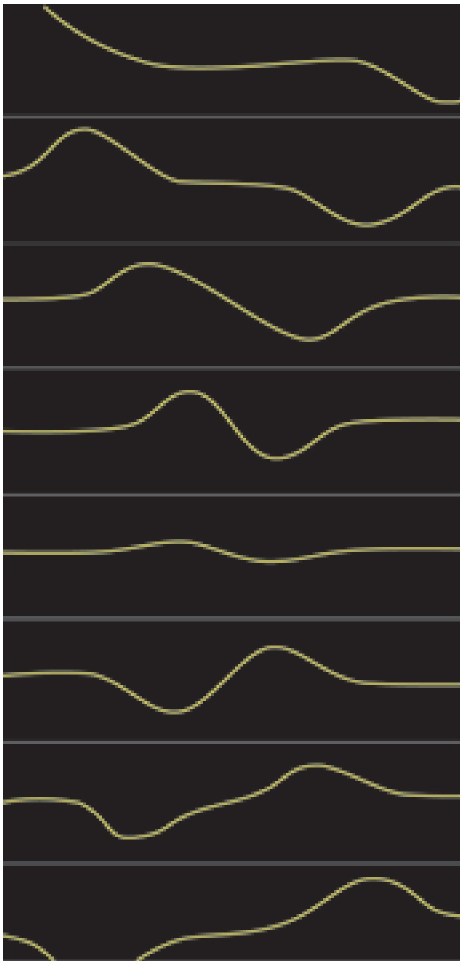


Figure 5.1
Two wave pulses carry their energy in opposite directions along the length of a spring with no net transfer of matter.

Water waves on the surface of water are approximate examples of transverse waves, as the particles vibrate at right angles to the direction of travel of the wave. This is seen in a floating cork or yacht bobbing up and down (vertically) as waves pass by horizontally.

Instead of sending a single pulse along the spring, the student could have continually oscillated their hand from left to right, effectively setting up a sequence of pulses or **wave train**. When the source of a disturbance undergoes continual oscillation, it will set up a **periodic wave** in a medium. Rather than experiencing a single disturbance as a pulse passes by, the particles of the medium will undergo continual vibration about a *mean position*. Many periodic waves in nature are *sinusoidal waves*, represented by the familiar sine and cosine graphs.

Figure 5.2
While waves moving along a spring or string are one dimensional, surface waves or ripples on water caused by dropping a pebble into a pond are two dimensional. Tiny speakers, for example, can send out sound waves in three dimensions.

The nature of sound

If you rest your fingers gently on your throat while speaking, you will feel the vibrations that are creating the sounds that make up your speech. The variety of vibrations you can feel are produced by the vocal cords in the larynx. Stand in front of a loudspeaker with the amplifier 'turned up loud' and you'll feel the same vibrations go right through your body. The process that produces these vibrations is complex, but a simple model on which we can base our initial observations is the tuning fork. When struck, a tuning fork produces a single note. A close look at the prongs will show that they are vibrating very rapidly. If the end of the vibrating tuning fork is dipped in water, drops of water will shoot out in different directions. Vibrations such as these are responsible for all the sounds we hear.

The transmission of sound

Can sound travel through a vacuum, such as outer space? Think back to the vibration of the tuning fork. As the prongs move back and forth, they hit air molecules, and these molecules hit the ones beyond them before rebounding to their original positions. In the meantime, the tuning fork will have moved back to the position where it will once again hit the air molecule, and the process is repeated. So each molecule in turn will vibrate back and forth about a mean position, receiving the motion from the tuning fork and passing it on to the molecules around it. But if there are no molecules, as in a vacuum, the vibration cannot be transferred.

Everyone is aware that sound travels in air and water. By ducking your head under water, you will hear that sound travels more easily in water than in air. For example, it is possible to hear a motorboat even when it is several kilometres away. Sound travels through water more rapidly and with less energy loss than through air. Many solids also transmit sound well. Someone who puts their ear on the railway tracks to hear the approach of a distant train utilises this fact.

SOUND is a form of mechanical energy transferred by the vibration of the molecules within the medium. Sound requires a medium in which to travel.

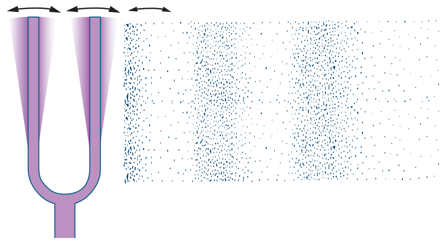


Figure 5.3
The rapid vibration of a tuning fork makes the air molecules around it vibrate in much the same way as the human larynx or a loudspeaker—creating the compressions and rarefactions of a pressure wave that we perceive as sound.

Physics file

Despite apparent evidence to the contrary from many science fiction epics, in space they cannot hear you scream. Space is largely a vacuum.

Figure 5.4
Loudspeakers convert electrical energy into sound by moving backwards and forwards to create the compressions and rarefactions of a sound wave.

Evidence for the wave nature of sound

If you want to throw a ball to a friend to catch, you should throw it in their direction. But if you want to 'throw' a sound to them, it does not matter if you are facing in a different direction—they will still hear you (although perhaps not as clearly). There does not even have to be a wall nearby for the sound to bounce off, because sound spreads out in all directions from the source. This kind of spreading behaviour is not possible with particle motion at everyday speeds, like that of a ball, but it is possible with waves. Consider, for example, the spread of ripples as a stone hits the surface of a pond.

Figure 5.5

The motion of a flame in front of a loudspeaker is clear evidence of the continuous movement of air to and fro as the loudspeaker creates sound.

Sound as longitudinal waves

A candle flame held in front of a large loudspeaker that is emitting a loud sound will move to and fro. This continuous movement indicates the direction of the vibrations in the air. Energy must be transferred from the loudspeaker to the air moving the flame, but by what means?

A stretched slinky spring can be used to visualise the movement of the air. A single pulse sent along the spring by moving it quickly backwards and forwards simulates the vibration. Figure 5.6 shows the formation of *compressions* in the spring that are produced when one end of the spring is pushed forwards, forcing a section of the coils closer together. The movement that follows, in the opposite direction, pulls the coils further apart and produces a *rarefaction*. Successive compressions and rarefactions move along the spring as the motion continues. In the spring:

- compressions occur in sections where the coils are closer together than average
- rarefactions occur in sections where the coils are further apart than average.

After a compression or rarefaction moves—or is *propagated*—along the spring, the spring returns to its original shape. Each part of the spring has been set in motion and the energy has travelled along the spring, but there has been no net movement of the spring itself. The series of compressions and rarefactions moving along a spring makes up a continuous wave. A wave such as this, in which the vibrations are in the same direction as the line of travel, is called a *longitudinal* wave.

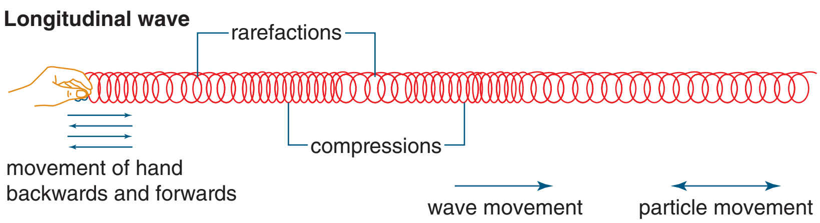
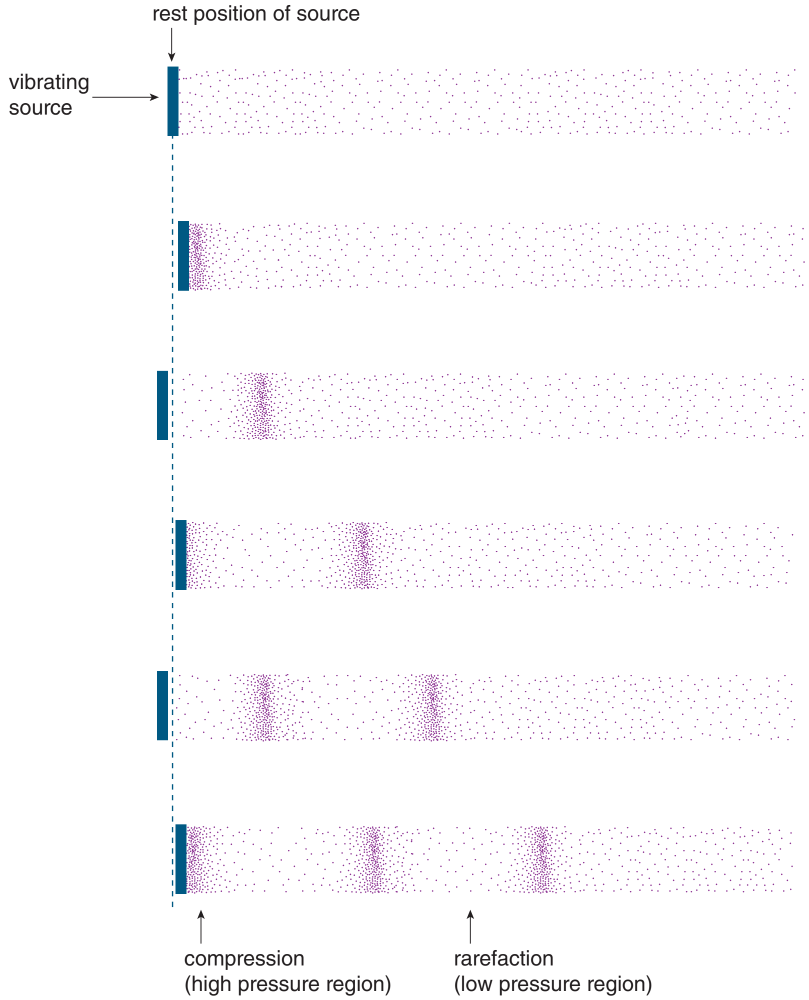


Figure 5.6

The transfer of energy in a spring by successive compressions and rarefactions, i.e. a longitudinal wave.

> All forms of ***WAVE MOTION*** cause a transfer of energy without a net transfer of matter.

The experiment with the candle flame indicates that the direction of the vibrations in air is along the direction that the sound is being propagated. Sound must, therefore, be a longitudinal wave, consisting of a series of compressions and rarefactions. The particles within the medium in which the sound is travelling, both atoms and molecules, are pushed closer together to form regions of increased air pressure (compressions), and pulled apart to form regions of lower air pressure (rarefactions). It is this periodic variation in air pressure that forms the sound wave.



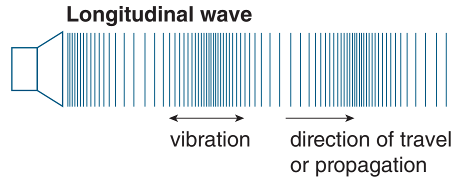


Figure 5.7
In a longitudinal wave, the vibrations of the individual particles are in the same direction as the direction of wave propagation.

Figure 5.8
In a medium in which there is no sound, the particles are evenly spread. As the source moves backwards and forwards, the continual vibration produces a series of compressions and rarefactions moving continuously away from the source. In this way, sound energy is propagated through a medium. Once the source ceases to vibrate, the particles in the medium return to approximately the same position. There is no net movement of the particles.

Transverse waves

A purely sideways motion of the particles forms another type of wave, often seen in vibrating strings, called a *transverse* wave. A slinky spring can also be used to demonstrate this form of wave motion (Figure 5.9). Crests and troughs replace the compressions and rarefactions of a longitudinal wave. In a transverse wave, the vibrations of the source that creates the wave are at right angles to the direction of travel of the wave. Once again, when the wave has passed, the particles in the medium will return to their original position: there is no net displacement of the particles.

Practical activities
1 Waves in a ripple tank
2 Waves in a slinky
3 Waves in a rope

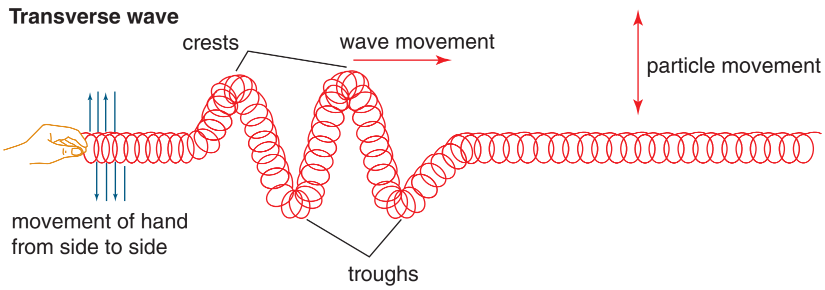


Figure 5.9
The transfer of energy in a spring by transverse movement.

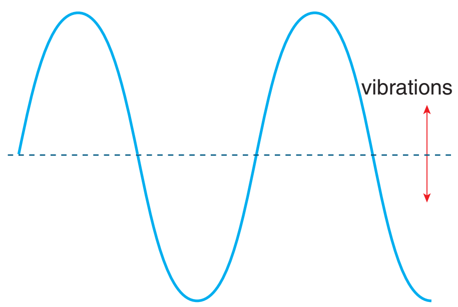


Figure 5.10

In a transverse wave, vibrations are at a right angle to the direction of travel of the wave.

Physics file

Keep in mind that the representation of a sound wave used throughout this section is one of pressure variation versus distance. The pressure variation refers to the change in pressure in the transition between compressions and rarefactions rather than the increase or decrease in pressure from normal atmospheric pressure associated with the positions of compressions and rarefactions themselves.

Physics file

The passage of sound through an elastic medium is actually by a combination of longitudinal and transverse waves. This can be best illustrated by watching a boat moving in deep water. You will find that its motion is both back and forth, as is the case for a longitudinal wave, and also obviously up and down, like a transverse wave. The resulting motion of the boat is elliptical. Try looking for this when watching surfers sitting on their boards.

Representing sound as waves

Picturing sound as a series of compressions and rarefactions in the medium in which it moves makes sense, but it is hard to show on paper. As a result, a series of conventions have been adopted that link the real situation with a pictorial representation.

Figure 5.11 shows a simple representation of a transverse wave. A wave like this one is called a *periodic* wave, because it regularly repeats. The *wavelength* of a periodic wave is the distance between successive points that have the same displacement from the rest position and where the particle is moving in the same direction. Points A and B in the diagram are such points; they are said to be *in phase*. The distance between them is the wavelength, denoted by λ (the Greek letter 'lambda') and usually measured in metres (m). Points C and D are also in phase with each other, but not with A or B. The point pair C and D is also one wavelength apart.

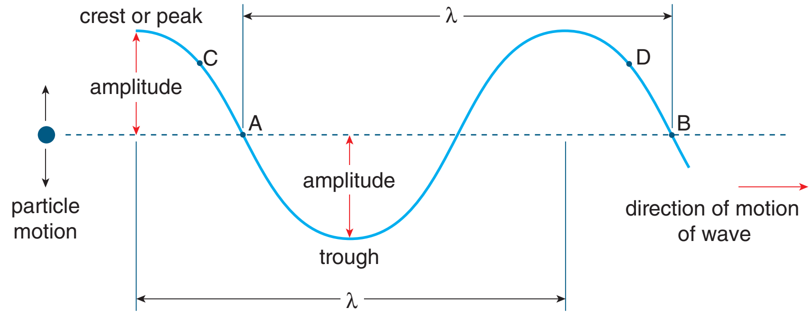


Figure 5.11

A simple pictorial representation of a transverse wave.

The amplitude, *A* (in metres), of a transverse wave is the distance from the rest position of a particle to the limit of a crest (positive maximum displacement) or trough (negative maximum displacement). The total distance from crest to trough is twice the amplitude. The frequency, *f*, of the wave is the number of waves (or cycles) that are repeated in 1 second, and is measured in cycles per second (s^{-1}), or hertz (Hz).

$$f = \frac{1}{T}$$

where T is the time taken for one cycle to be completed in seconds (s) and f is the frequency of the wave in hertz (Hz) or cycles per second (s^{-1}).

For example, if 10 crests pass a given point in 1 second, then the frequency is 10.0 cycles per second, or 10.0 Hz. The period, or time for each complete wave, will be 0.100 seconds.

Since these definitions apply to all waves, let's examine how they apply to longitudinal sound waves. In Figure 5.12a, W and X are points that experience the same change in pressure from the rest position, so they are one wavelength apart. Y and Z are also in phase and one wavelength apart. This longitudinal wave can be drawn in a similar simple style to the representation of a transverse wave shown in Figure 5.11. A graph of pressure variation against distance from the source at a particular instant in time can be used to represent the periodic changes from compression to rarefaction. The result is a graph of the wave like that shown in Figure 5.12b. W and X are now clearly one wavelength apart. The frequency, f, is the number of complete cycles per second, and the period, T, is the time for a particle to complete one cycle from being at a point of compression to rarefaction and back again. The larger the amplitude of the pressure variation, the greater the wave energy that is needed to produce it, so a louder sound will be represented by a greater amplitude.

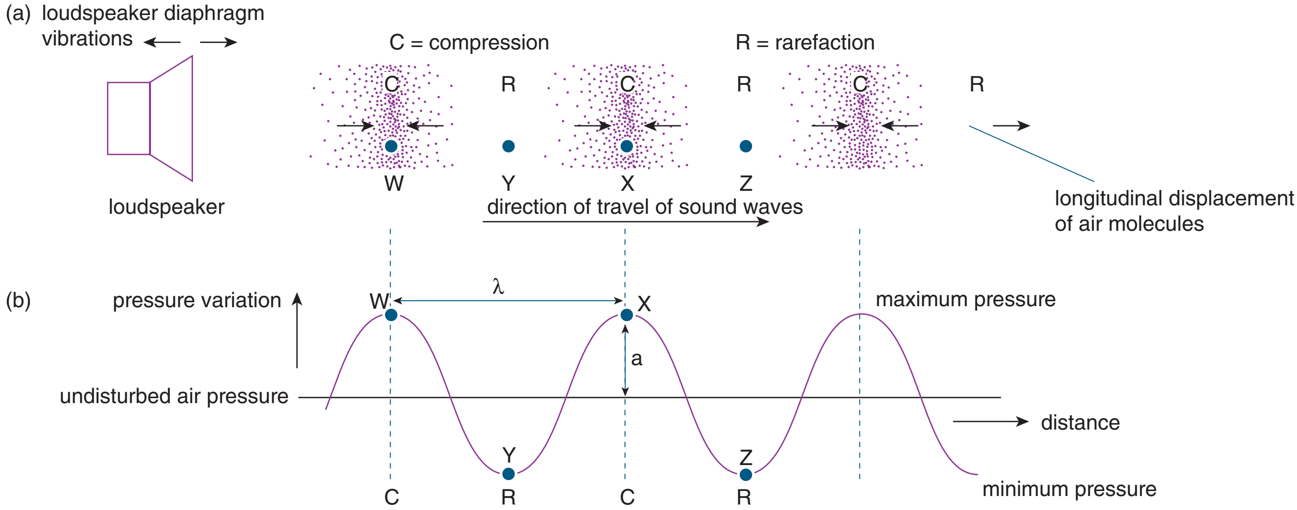
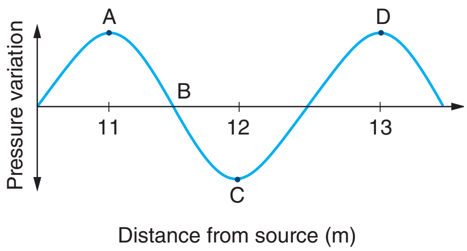


Figure 5.12
A longitudinal wave as a simple pictorial representation.

✓ Worked Example 5.1A

A note sung by a school choir is transmitted through air as a longitudinal wave. The graph below represents the pressure variation of the sound wave at a particular instant in time.



a What is the wavelength of the sound wave?
b Which of the points A–D represent compressions?

Solution

a Points A and D represent particles currently at points of the same pressure variation. They are one wavelength apart. Using the scale shown on the graph:

$\lambda = 13 - 11$
$= 2$ m

b Points of compression are represented as points of maximum positive pressure; hence, A and D are currently at points of compression.

The wave equation

Mechanical waves in different media will propagate at different speeds. Since it is the physical transfer of vibrations that occurs as the wave travels, the properties of the medium, such as its density and temperature, largely determine this speed. Since velocity = displacement/time, the velocity of a wave can be determined by examining a single pulse and measuring the time taken for it to travel a specific distance. If the distance used was equal to one wavelength, λ, then by definition the time taken must be one period, *T*. Hence:

Interactive tutorial
The wave equations

THE WAVE EQUATION is given by:

$$\boldsymbol{v} = \frac{\lambda}{T} \quad \text{and} \quad T = \frac{1}{f}$$

$$\therefore \boldsymbol{v} = f\lambda$$

where $\boldsymbol{v}$ is the velocity of the wave (m s^{-1}), *f* is the frequency (Hz) and λ is the wavelength (m).

The wave equation applies to all types of wave motion. As each wave is created, it is carried away from the source at a particular velocity determined by the medium. Therefore, the waves will have travelled a certain distance before the next wave is created. Together, this velocity and the frequency of the **wave source** will determine the resulting wavelength. For a source of a given frequency, a fast medium would result in a longer wavelength, a slow medium would result in a shorter wavelength. In a given uniform medium the velocity of a wave will be constant. In this case the wavelength and frequency would be inversely proportional to one another. High-frequency sources would create short wavelengths and vice versa.

✓ Worked Example 5.1B

Waves are created on the surface of the water in a wave pool by a device that dips up and down 20.0 times per minute. The velocity of the resultant wave is 3.00 m s^{-1}.

a State the frequency, period and wavelength of the observed wave.

b What happens to the value of the velocity and wavelength if the frequency of the source is doubled?

Solution

a Since frequency is defined as the number of waves or cycles that pass a given point per second:

f = 20.0 per minute

$$f = \frac{20.0}{60} = 0.333 \text{ Hz}$$

f = 0.333 Hz

$$T = \frac{1}{f} = \frac{1}{0.333} = 3.00 \text{ s}$$

The period is therefore 3.00 s

The wavelength is then:

f = 0.333 Hz

$\boldsymbol{v}$ = 3.00 m s^{-1}

$$\boldsymbol{v} = f\lambda$$

$$\lambda = \frac{\boldsymbol{v}}{f} = \frac{3.00}{0.333} = 9.00 \text{ m}$$

b The velocity is determined by the properties of the medium and so it is unchanged, $\boldsymbol{v}$ = 3.00 m s^{-1}. At a set velocity, doubling the frequency will result in wavelengths of half the original value, i.e. 4.50 m.

5.1 SUMMARY Wave properties

- All waves involve energy being transferred from one location to another, without any net transfer of matter.
- Periodic waves are characterised by several features—frequency, period, wavelength and amplitude.
- The frequency, *f*, of a wave is defined as the number of waves or cycles that pass a given point per second; it is measured in cycles per second (s^{-1}) or hertz (Hz).
- The period, *T*, of a wave is the time taken for one cycle to be completed; it is measured in seconds.
- The amplitude, *A*, of a wave is the value of the maximum displacement of a particle from its mean position.
- Wavelength is defined as the minimum distance between two points in a wave that are in phase. Alternatively, wavelength can be defined as the distance that a wave travels during one period. It is measured in metres (m).
- The wave equation states that $v = f\lambda$, where v is the velocity (m s^{-1}), f is the frequency (Hz) and λ is the wavelength (m).
- Sound is a longitudinal wave. Vibrations occur in the same direction as the direction of propagation of the wave.
- Sound waves are made up of a periodic series of compressions and rarefactions representing changes in pressure within the medium.
- A compression is a region of higher than average air pressure. A rarefaction is an area of lower than average air pressure.
- Transverse waves are produced by particles which vibrate at right angles to the direction of travel of the wave (e.g. waves in a piece of rope).

5.1 Questions

1 What do mechanical and electromagnetic waves have in common?

2 Suggest why mechanical waves generally travel faster in solids than in gases.

3 A guitar string is plucked and it undergoes 40 vibrations in 0.250 s. What is the frequency and period of the sound wave produced?

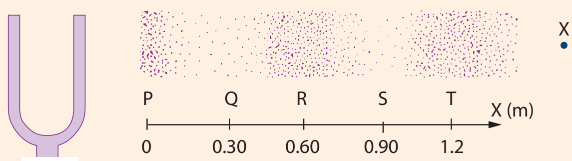
4 A student sending transverse waves along a slinky spring completes three full cycles in 1.00 s. The wave train produced has a total length of 1.80 m. Determine the speed of the wave in the spring.

5 A wave generator in a ripple tank produces eight vibrations per second. If you wanted to produce waves with a greater wavelength, would you increase or decrease the frequency of the generator? Explain.

6 Which one of the following correctly describes the reason we can hear sound from a vibrating tuning fork?
- A The air molecules adjacent to the prongs are propelled towards our ears and eventually strike the eardrum.
- B The kinetic energy of the vibrating prongs is transmitted instantaneously to the eardrum in the same way that electromagnetic energy (or light) is transmitted.
- C The prongs produce vibrations in adjacent air molecules, which results in the transfer of energy to other adjacent molecules and eventually to the eardrum.

7 Which one or more of the following statements is true?
- A Sound can travel in a vacuum.
- B During the transmission of sound, air molecules are permanently displaced from their original positions.
- C Sound exhibits wave-like behaviour.
- D The transfer of sound energy is independent of the medium.
- E The transmission of sound involves only the transfer of energy and not of matter.

8 Which of the following describes the direction of the vibrations of the air molecules in the following wave?



A ↔ B → C ↓ D ↑ E ←

9 For the wave in Question 8, which of the following alternatives correctly describes the direction of energy transfer of the sound between the tuning fork and point X?
A ↔ B → C ↓ D ↑ E ←

10 A microphone is connected to a digitiser and placed at point X in the diagram in Question 8. The pressure variation vs. time trace displayed on the connected computer screen is as follows.
- a Describe the type of energy transformation occurring at the microphone.
- b At what time, or times, did maximum pressure variation occur?

5.2 Wave behaviour

In the first section of this chapter, we established the concept of a wave by considering how waves are produced and transmitted. We will now look at three particular phenomena of wave behaviour: reflection, refraction and diffraction.

Representing waves

In order to clearly represent waves, and the direction in which they travel, we will use two models. The wavefront model represents waves as a line joining neighbouring particles of maximum amplitude; this line represents the crest of the wave and will be perpendicular to the direction of energy flow in the wave. The wave model is particularly useful for representing a change in wavelength or the presence of waves in a region, but is not as clear when representing a change in direction. The ray model represents waves by indicating with an arrow the direction in which their energy is moving. This model is particularly useful when indicating a change in the direction of the wave.

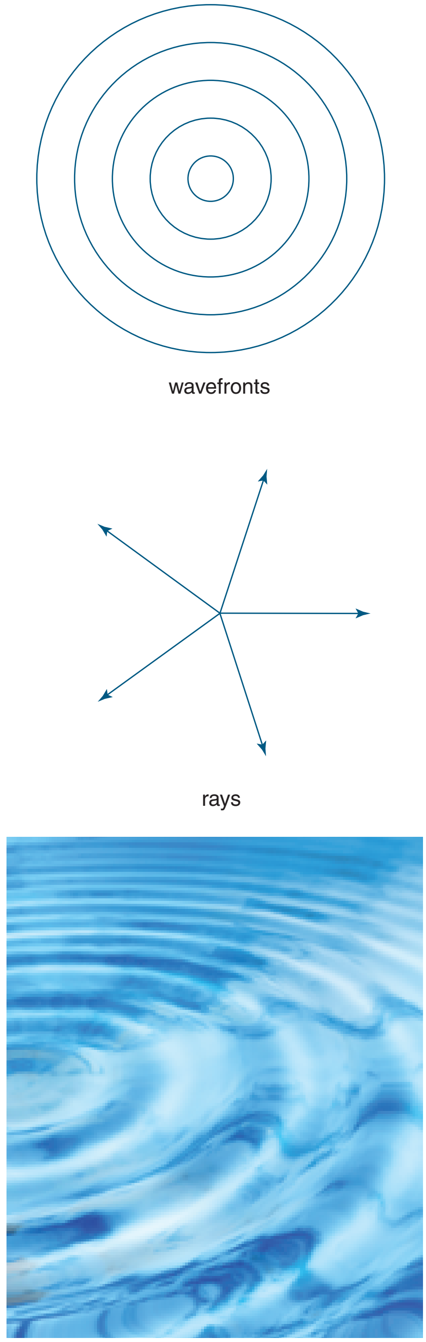


Figure 5.13
Two methods of representing the same water waves.

Reflection of waves

Echoes provide the most obvious evidence that sound waves are reflected. Like all waves, sound can be reflected when it strikes an obstacle and will be at least partially reflected when moving from one medium to another. A simple arrangement of a directional source and microphone can be used to confirm that sound obeys the same law of reflection that light obeys (Figure 5.14). A directional source can be made by placing a loudspeaker, connected to a signal generator, inside a tube. A microphone placed inside a tube and connected to a cathode ray oscilloscope (CRO) or digitiser makes a simple directional director. The CRO allows the strength of a reflected sound wave to be measured on a small screen with a grid printed on it.

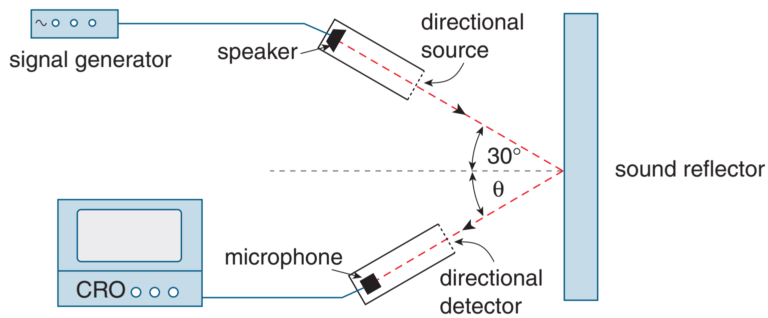


Figure 5.14
Simple directional equipment can be used to confirm that, as with other forms of waves, the angle of incidence equals the angle of reflection. As is the case with light, a concave surface will focus sound waves and a convex surface will cause sound to disperse.

Practical activities

4 Reflection of waves in a ripple tank
5 Reflection in a plane mirror

Using this equipment, you would quickly discover that the maximum reflected signal strength occurs when the angle of incidence is equal to the angle of reflection. Sound obeys the law of reflection. In addition, when the sound is reflected in the same medium there is no change in its frequency, wavelength or speed.

If the reflecting surfaces are curved, as shown in Figure 5.15b and c, then sound waves will converge or diverge, like light. Simple ray tracing can determine the path the sound will follow.

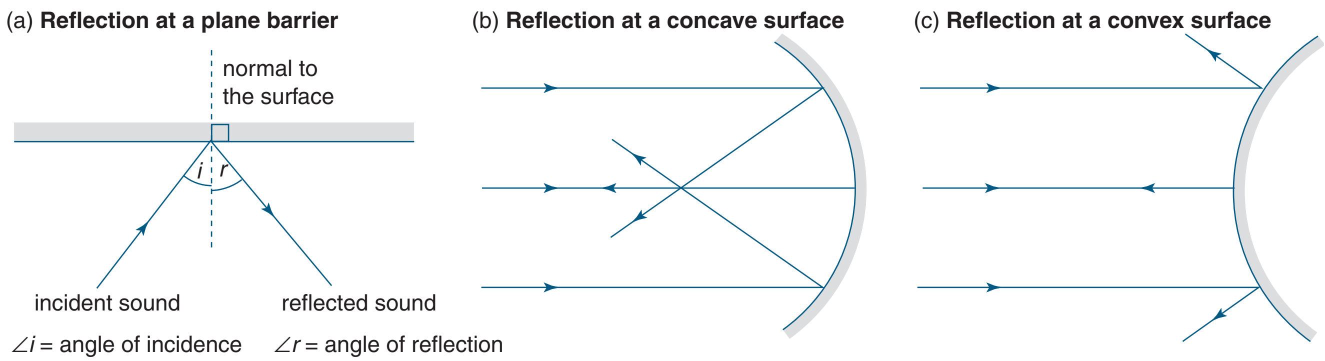


Figure 5.15
Ray tracings of the paths of reflected sound waves.

The **_LAW OF REFLECTION_** for waves states that: the:

- angle of incidence (i) = angle of reflection (r) for a wave
- incident ray, normal and reflected ray are co-planar
- incident ray and reflected ray are on the opposite sides of the normal.

Physics file

The phase change of a wave on reflection from a fixed end can be explained in terms of Newton's third law of motion. When the pulse arrives at the fixture, the rope exerts a force on the fixture. The fixture exerts an equal and opposite force on the rope. This produces a pulse that is in the opposite direction to the original pulse; that is, a change in phase has occurred.

Waves meeting barriers

At the boundary of the medium, the energy that was being carried by the wave may undergo different processes. Some of the energy may be *absorbed* by or *transmitted* into a new medium, and some energy may be *reflected*.

The extent to which these processes occur depends on the properties of the boundary. We shall examine the case of a transverse wave pulse travelling in a heavy rope that has one end tied to a wall. As shown in Figure 5.16a the wave travels to the boundary and we can see that it is reflected with almost no energy loss since the original amplitude is maintained. The wave, however, has been inverted; this can also be described as a *reversal in phase*. (The definition of *phase* was discussed in the previous section.) Since a crest would reflect as a trough and a trough would reflect as a crest, we can say that the phase of the wave has been shifted by $\frac{1}{2}\lambda$.

A **_WAVE REFLECTING FROM THE FIXED END_** of a string will undergo a phase reversal; that is, a phase shift of $\frac{\lambda}{2}$.

Now consider the situation in which the end of the rope is free to move. As shown in Figure 5.16b, the wave travels to the end of the rope and we can see that it is reflected with *no* reversal in phase. Since a crest would reflect as a crest and a trough would reflect as a trough, we can say that there was no change of phase.

A **_WAVE REFLECTING FROM THE FREE END_** of a string will *not* undergo a phase reversal.

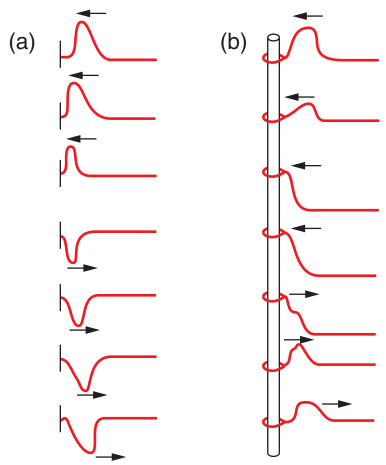


Figure 5.16
(a) The reflection of a wave at an unyielding boundary produces a phase shift of ½λ. Note that otherwise the shape of the wave is unaltered.
(b) The reflection of a wave at a free-end boundary does not produce a phase shift.

Physics in action — Reflections NOT wanted!

The stealth aircraft is designed so that its body is as poor a reflector as possible. The main way in which a passing aircraft is detected by others is with the use of radar. A radar transmitter sends out pulses of radio waves or microwaves and a receiver checks for any reflections from passing aircraft. By analysing the reflections, radar systems can work out the position, speed and perhaps even the identity of the passing aircraft.

Stealth aircraft are designed to create as little reflection of these waves as possible. The shape of the stealth aircraft is the most important factor. It does not have any large vertical panels on the fuselage that would act like mirrors, nor a large vertical tail. It has no externally mounted devices such as missiles or bombs. It does not include any surfaces that meet at right angles. These would act like the corners in a billiard table and bounce the waves right back to their source. Instead curved surfaces on the stealth aircraft are designed to reflect waves sideways or upward wherever possible. A thick coat of special paint that absorbs radio waves is used on its surface. Although not completely undetectable, with the right shape and coating a large stealth plane can produce the same amount of wave reflection as an average sized marble!

Figure 5.17
The Lockheed Martin F-35 Joint Strike Fighter is a long-range stealth fighter.

Refraction of waves

Refraction is associated with all types of waves. Refraction is the bending of waves when waves travel from one medium into another medium. When sound travels from one medium to another, the direction of the sound can change. This is always associated with a change in the speed of the sound. The speed of sound can be quite different in various media since the speed of sound depends on a number of factors including density, elasticity and temperature. Sound travels slower in media of higher density, faster in hotter media, particularly gases, and faster in more elastic or 'stiffer' media.

Sound, like any wave, will bend towards the normal to the interface between two media if it is slower in the second medium. It will bend away from the normal if it travels faster in the second medium.

When sound refracts into a new medium, the frequency does not change. The frequency of a periodic wave is a fundamental property of the wave and does not change when the wave reflects, refracts or, as will be seen later, diffracts. Since the speed of sound changes when the sound travels into a new medium and the frequency remains unchanged, the wavelength must also change. This makes sense if you consider the wave equation: $v = f\lambda$. The speed of a wave is directly proportional to its wavelength if the frequency is to remain unchanged.

Physics file

A sound modelling system designed by Bose uses computer programs to model the acoustic behaviour of architectural designs. It was used in the design of Tottenham Hotspur's football stadium in the UK. The system takes into account the materials used in the construction of buildings and the shape of the building. It also allows for temperature and humidity variations in the building.

Interactive tutorial
Refraction

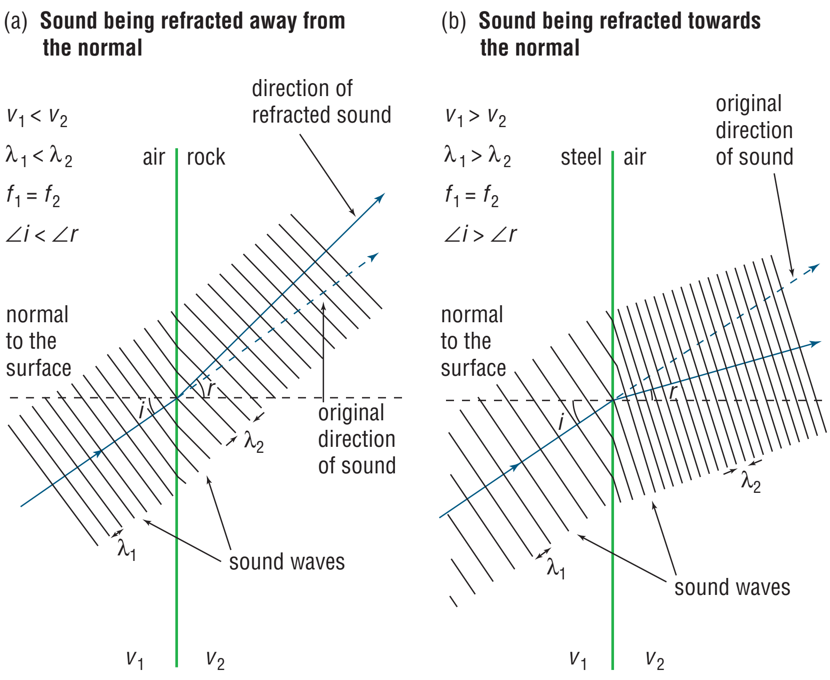


Figure 5.18

Sound waves being refracted at a boundary. *i* is the angle of incidence and *r* is the angle of refraction. The wavelength of the waves in the first material is λ_1 and in the second material it is λ_2. The velocity of the waves in the first material is v_1 and in the second material is v_2. The frequency of the sound is the same in both materials. (a) If sound waves from one medium enter another where the speed of sound is faster, they will be refracted away from the normal to the surface and their wavelength will be increased. (b) If sound waves from one medium enter another where the speed of sound is slower, they will be refracted towards the normal to the surface and their wavelength will be reduced.

Figure 5.18 shows the details of what happens when sound transmits into new media. The relationship between angles of incidence, angle of refraction, wave speeds and wavelengths is known as Snell's law. It is named after Willebrord Snell, a Dutch scientist who studied the refraction of light and discovered the law of refraction around 1621.

SNELL'S LAW is:

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2} = \frac{\lambda_1}{\lambda_2}$$

where *i* and *r* are the incident angle and the refracted angle, v_1 and v_2 are the speeds of sound in the first and the second medium and λ_1 and λ_2 are the wavelengths in the first and the second medium.

The ratios in Snell's law are sometimes referred to as the relative refractive index of the two media, but this term is usually used when referring to refraction of light rather than refraction of sound.

Air of different temperatures can be considered as two different media and sound will change direction as it moves from cold air into warm air.

Physics file

The speed of sound is affected by two bulk properties of matter: 'stiffness' and 'inertia'. Inertia is measured by the density of the medium. The more dense a medium, the slower will be the propagation of sound. This is because sound propagation involves vibrations in the medium. More mass per unit volume makes it more difficult for a given force to move the medium so the vibrations take more time. Many people wrongly believe that more dense media propagate sound faster because sound propagates faster in solids than in gases and they think that density is the only factor involved. The stiffness of solids is vastly greater than the stiffness of gases and this is the reason that sound travels faster in solids than in gases.
Stiffness is more correctly called elasticity. Elasticity is the tendency of a medium to return to its original shape after it has been bent, stretched or compressed. It is the elasticity of a medium that causes sound to propagate throughout the medium.
The speed of sound depends very simply on stiffness and density as follows:

$$v = \sqrt{\frac{E}{\rho}}$$

where *v* is the speed of sound (m s^{-1}), *E* is Young's modulus (N m^{-2}) and ρ is the density of the medium (kg m^{-3}).
Temperature affects the speed of sound in air because changes in temperature change the density of the air without changing its elasticity. Higher temperatures cause the air to expand and become less dense. Lower density results in a greater sound speed.

Practical activities

7 Refraction of continuous water waves
8 Investigating refraction: Snell's law

✓ Worked Example 5.2A

A 500 Hz sound from a rock concert travels from cold air where the speed of sound is 330 m s^{-1} into warmer air where the speed of sound is 355 m s^{-1} and strikes the boundary between the two regions of air with an angle of incidence of 30°.

a Find the wavelength in each region of air.

b Find the angle of refraction into the second region.

Solution

a Using the wave equation:

$\lambda_{cold} = 330 \text{ m s}^{-1}$

$f = 500 \text{ Hz}$

$$v = f\lambda$$

$$\lambda_{cold} = \frac{v}{f} = \frac{330}{500} = 0.660 \text{ m}$$

$\lambda_{warm} = 355 \text{ m s}^{-1}$

$f = 500 \text{ Hz}$

$$v = f\lambda$$

$$\lambda_{warm} = \frac{v}{f} = \frac{355}{500} = 0.71 \text{ m}$$

b Using Snell's law:

$i = 30.0°$

$v_{cold} = 330 \text{ m s}^{-1}$

$v_{warm} = 355 \text{ m s}^{-1}$

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2}$$

$$\sin r = \frac{v_2 \sin i}{v_1} = \frac{355 \sin 30.0°}{330}$$

$$\sin r = 0.5379$$

$$r = 32.5°$$

Total internal reflection

Whenever refraction occurs a proportion of the wave energy is reflected. This happens when sound refracts into a new medium. While most of the energy may transmit into a new medium, some is reflected. When sound refracts into a medium in which the speed is greater, it is possible for total internal reflection to occur. The minimum angle of incidence at which this will occur is called the critical angle.

Mathematically, this situation can be recognised if Snell's law gives sin r a value greater than one. Alternatively, the maximum possible value for the incident angle can be calculated for any two media. The critical angle is the angle of incidence for which the angle of refraction would be 90°.

Practical activities

10 Total internal reflection in prisms

11 Optical fibre bend loss

12 Fibre optic cladding

✓ Worked Example 5.2B

At what minimum incident angle would sound need to strike water from air if it is to reflect completely? The speed of sound in air is 344 m s^{-1} and in water it is 1500 m s^{-1}.

Solution

In this example, the critical angle needs to be calculated using Snell's law:

$r = 90.0°$

$v_{water} = 1500 \text{ m s}^{-1}$

$v_{air} = 344 \text{ m s}^{-1}$

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2}$$

$$\sin i = \frac{v_1 \sin r}{v_2} = \frac{344 \sin 90.0°}{1500}$$

$$\sin r = 0.0229$$

$$r = 13.3°$$

In general, the ***CRITICAL ANGLE*** is given by:

$$\sin c = \frac{v_1}{v_2}$$

where c is the critical angle, v_1 is the speed of sound in the first medium (m s^{-1}) and v_2 is the speed of sound in the second medium (m s^{-1}).

The critical angle is quite small if there is a large difference between the speed of sound in the first medium and the speed of sound in the second medium. In the worked example on the previous page, the speed of sound in the second medium was about five times the speed of sound in the first medium, giving a critical angle of only 13.3°. This means that sound will only transmit into water if the angle of incidence is less than 13.3° and otherwise will just reflect off the surface. A similar calculation for sound going from air into steel, where the speed of sound is 5000 m s^{-1} reveals that the critical angle is only about 4°. This explains why liquids and solids are good reflectors of sound. The sound can only refract into a solid or a liquid if it meets the surface at a very restricted range of incident angles. All other sound meeting the interface between air and a solid or a liquid will be reflected back into the air. Figure 5.19 shows this situation.

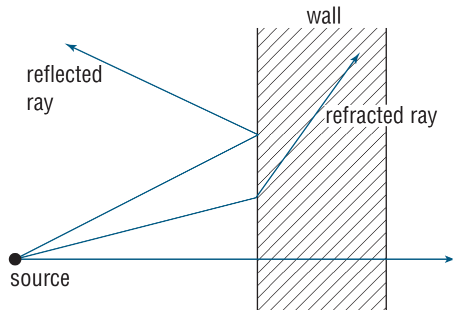


Figure 5.19
Sound transmitted through air towards a wall will only refract into the wall at a narrow range of incident angles. If the critical angle is only, say, 10° then sound incident on the wall at any angle larger than 10° will be reflected rather than refracted into the wall. This is one of the reasons that solid walls are good sound insulators. They are good reflectors of sound.

Diffraction of sound

If you hide from someone around a corner of a building, they can still hear you if you make a sound, even if there are no reflecting surfaces nearby. This well-known ability of sound to travel around corners provides further evidence that sound is wave-like in nature. Reflection alone cannot account for all the indirect sounds. Another clue is that higher frequency sounds can be heard more clearly if the listener is directly in front of the source, while lower frequencies can be heard quite clearly from a wide range of angles. This has major implications for the design of sound reproduction systems and comes about as a result of a wave phenomenon known as *diffraction*.

How much a particular wave spreads on passing an obstacle or through an aperture will depend upon its wavelength in relation to the size of the obstacle or aperture. Sound waves passing through an aperture or past an obstacle that is larger than the wavelength will not be significantly diffracted, but apertures or obstacles that are comparable to the wavelength or smaller will cause considerable diffraction, and the sound will spread out.

Practical activities

13 Diffraction of continuous water waves
14 Diffraction of light

DIFFRACTION is the bending of waves as they pass the edge (or edges) of an obstacle or pass through an aperture.

Diffraction and wavelength

As a general rule, the amount of diffraction will depend on the ratio of the wavelength, λ, of the sound to the width, w, of the aperture or obstacle.

SIGNIFICANT DIFFRACTION will occur when the wavelength is of at least the same order of magnitude as the width of the obstacle or aperture.

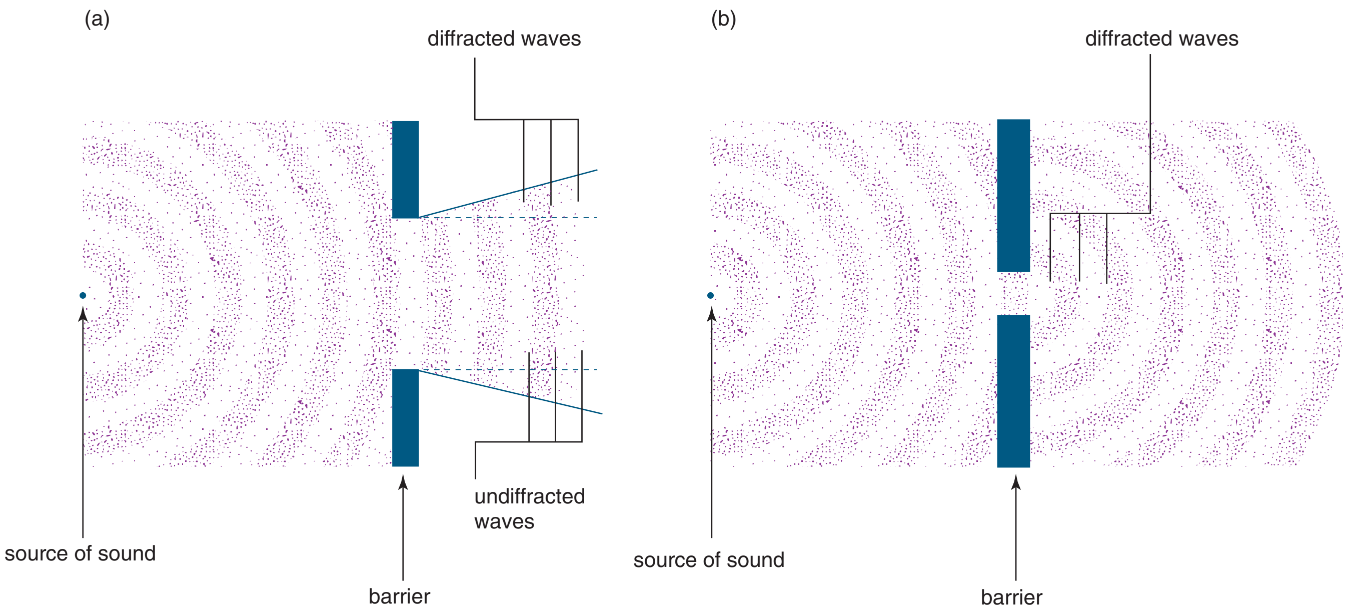


Figure 5.20

The amount of diffraction depends on the size of the gap, or aperture, and the wavelength. (a) When the aperture is much larger than the wavelength, only limited diffraction will occur. (b) Wavelengths larger than the aperture width will result in significant diffraction.

When the wavelength is small, obstacles will cast greater sound 'shadows' (i.e. regions of no disturbance) and waves will spread out less. For the same reason, long-wavelength sounds spread out to fill a space, making it difficult to determine the exact source of the sound.

Diffraction and frequency

Earlier we saw that in a given medium at a constant temperature, the wavelength of a sound wave is inversely proportional to its frequency ($\lambda \propto \frac{1}{f}$). The speed of sound does not depend on either f or λ. This means that shorter wavelengths have higher frequencies, and longer wavelengths have lower frequencies. The wavelengths of sounds within the normal human hearing range are between about 2 cm and 20 m. A typical human voice has a wavelength of around 1 m, so voices diffract easily through doorways and around large obstacles.

Higher frequencies are diffracted less, so they are more directional; that is, it is easier to hear them from a particular direction, and they may not be heard as easily from other directions. Ultrasound (with frequencies greater than 20 000 Hz) is used for sonar and ultrasonic motion detectors because its shorter wavelengths mean that diffraction is very limited, so the sonar beam tends to travel directly to and from an object with only a small degree of spread.

Lower frequencies with larger wavelengths are diffracted to a greater extent, and low-frequency sound will readily fill a room. The source may often be difficult to locate by sound direction alone. Sub-woofer speakers of audio systems are designed around this idea. The very low-frequency sounds they produce have long wavelengths, so they are easily diffracted and hence appear to come from all around the room. Hence, one low-frequency loudspeaker is adequate. A true stereo effect cannot be detected at low frequency: this effect depends mainly on the higher, more directional frequencies produced from mid-range and 'tweeter' speakers, where two or more separated loudspeaker systems are used.

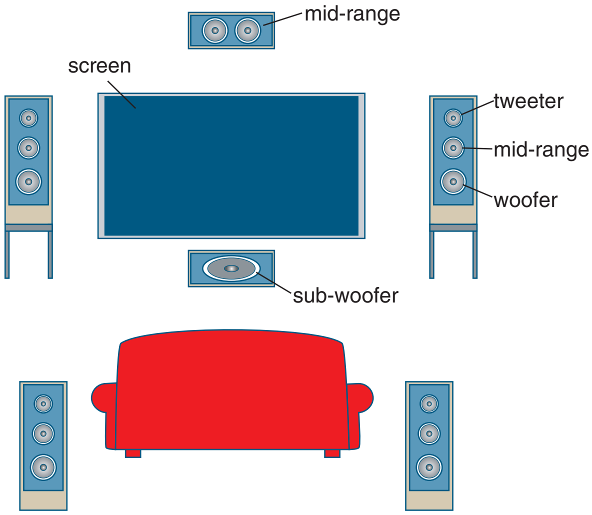


Figure 5.21
Most modern speaker systems use moving coil loudspeakers designed to reproduce particular ranges of frequencies: large bass 'woofers', or even a single sub-woofer, for the lower non-directional frequencies (20–500 Hz), multiple mid-range units (500 Hz – 4 kHz) and many directional 'tweeters' for the high frequencies (4–20 kHz). The ability of the system to authentically reproduce sounds often depends more on the installation than on the speaker system itself.

✓ Worked Example 5.2C

When sound waves of a high frequency, for example 9000 Hz, strike an obstacle such as a person's head, they leave a distinct sound shadow in which little of the sound can be heard. If one ear is closer to the sound source than the other, these higher frequencies will be heard as louder by the ear closer to the source.

a Assuming that the speed of the sound is 340 m s^{-1}, calculate the wavelength of a 9000 Hz sound.

b Explain why this high-frequency sound leaves a sound shadow on one side of a person's head.

Solution

a Using the wave equation:

$v = 340$ m s^{-1}

$f = 9000$ Hz

$$v = f\lambda$$

$$\lambda = \frac{v}{f} = \frac{340}{9000}$$

$$= 3.78 \times 10^{-2} \text{ m}$$

Hence, the wavelength is 0.0378 m or 3.78 cm.

b A person's head is about 20 cm in diameter. The wavelength of a 9000 Hz sound (3.78 cm) is significantly smaller than this, and since diffraction is only significant when $\frac{\lambda}{w} \approx 1$ or more, diffraction will be minimal and the sound will not bend significantly around the head.

✓ Worked Example 5.2D

A flute and a tuba are being played at the same time. The flute is producing a note with a frequency of 2000 Hz and the tuba is producing a note with a frequency of 125 Hz at the same volume. A listener to the side of the auditorium complains that the tuba is drowning out the flute. How can this be so?

Solution

The higher-frequency sound of the flute corresponds to a shorter wavelength, so it will be diffracted less. Thus it will be more directional than the sound of the tuba and it will not be heard as well at the sides of the auditorium.

5.2 SUMMARY Wave behaviour

- Reflecting sound waves obey the law of reflection, i.e. angle of incidence i = angle of reflection r.
- Like other waveforms, sound can be focused by a concave surface or spread by a convex surface. Simple ray tracing will determine the path the waves will follow.
- Refraction refers to the change in direction of a wave when sound moves from one medium to another and changes its speed of propagation.
- Frequency remains unchanged when sound moves to a new medium. Its speed and wavelength may change.
- When sound moves to a medium in which its speed is greater, its direction of propagation bends away from the normal. When sound moves to a medium in which its speed is lower, its pathway bends towards the normal.
- The relationship between incident angle, refracted angle, the speeds in the media and the wavelengths in the media is given by Snell's law:

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2} = \frac{\lambda_1}{\lambda_2}$$

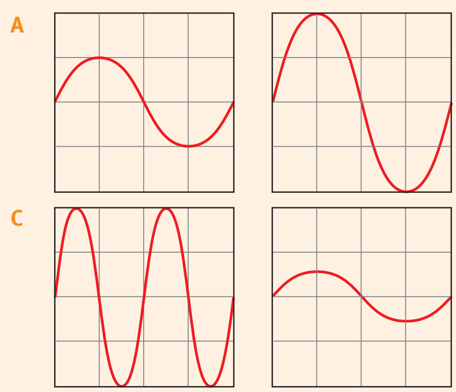
- If the angle of incidence is greater than the 'critical angle' for sound moving into a medium where its velocity is greater, then the sound will be totally reflected and will not enter the new medium.
- Diffraction is the bending of waves around the edge (or edges) of a barrier or aperture. The fact that sound diffracts provides further evidence that sound is wave-like in nature.
- The amount of diffraction depends on the ratio of the wavelength of the sound (λ) to the width of the opening or obstacle (w). Significant diffraction occurs when the wavelength is at least the same order of magnitude as the width of the opening or obstacle, i.e. when $\frac{\lambda}{w} \geq 1$.
- Higher frequency sounds have shorter wavelengths. As a result, they will diffract less than lower frequency sounds, which have longer wavelengths, and they will be more directional.

5.2 Questions

The following information relates to questions 1–4.

A directional sound source set at a constant angle of 30° to the horizontal bounces sound waves off a reflector, as shown in Figure 5.14. A directional detector, whose angle θ to the horizontal can be varied, allows the reflected signal to be displayed on a cathode ray oscilloscope. The figure at right shows four different displays produced, A–D. The oscilloscope's screen is graduated in centimetres and the time base is set at a constant value of 0.25 ms cm^{-1}. The speed of sound in the room is 340 m s^{-1}.

Initially the frequency of the signal generator was set at 1.00 kHz and the amplitude at R. The corresponding display for θ = 30.0° was A.



1 a What is the wavelength of the incident sound wave?
b What is the wavelength of the reflected sound wave for $\theta = 30.0°$?
c What is the speed of the reflected wave for $\theta = 30.0°$?

2 Which one of the displays A–D could correspond to $\theta = 40.0°$ for this frequency and amplitude? Justify your answer.

3 The frequency of the signal generator is now increased to 2.00 kHz, and the amplitude is increased to 2A.
a Calculate the wavelength of the incident wave.
b What is the wavelength of the reflected wave for $\theta = 30.0°$?
c What is the speed of the reflected wave for $\theta = 30.0°$?

4 Which one of the displays A–D represents the reflected sound wave for $\theta = 30.0°$ for this frequency and amplitude?

The following information relates to questions 5–8.

The velocity of sound in air increases by about 0.60 m s^{-1} for each degree of increase in temperature. The speed of sound in air at 0.0°C is 331 m s^{-1}. A flute plays a note of 880 Hz near the ground where the temperature is 20.0°C. A short distance above the ground there is a rapid change in the temperature into a layer of air at 30.0°C.

5 What is the speed of sound and the wavelength of the note in the air near the ground?

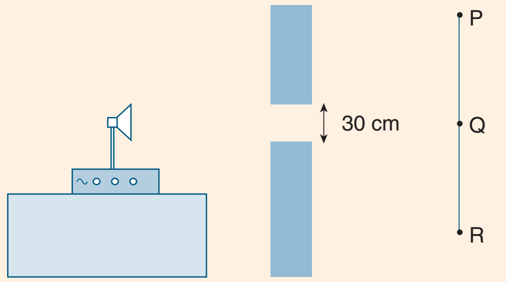
6 What is the speed of sound and the wavelength of the note in the warm layer of air?

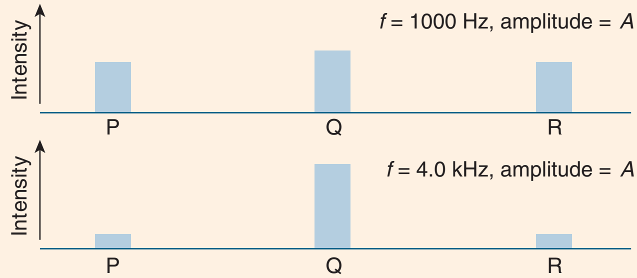
7 If the sound is incident on the warm layer at an angle of 50°C, what will the angle of refraction be in the warm air?

8 At what angle must the sound be incident on the warm air layer for the sound to be reflected rather than refracted?

The following information relates to questions 9 and 10.

A loudspeaker connected to a signal generator directs sound waves at a soundproof barrier in which there is an opening 30 cm wide, as shown in the following diagram. The computer-generated intensity graphs of the sound detected at points P, Q and R are shown. The speed of sound in both sections has a constant value of 340 m s^{-1}.





Initially the frequency of the wave generator is set at $f = 1000$ Hz, and amplitude = A.

9 a Calculate the wavelength of the sound produced.
b Explain why noticeable sound intensity levels can be detected at points P and R.

The frequency of the signal generator is now set at $f = 4.0$ kHz, amplitude = A.

10 a What is the wavelength of this sound?
b Explain why the intensity levels at points P and R are significantly less at this frequency than at 1000 Hz.
c Why is the intensity of sound at point Q higher at this frequency?

5.3 Wave interactions

The sounds produced by acoustic musical instruments and the human voice are the product of the interaction between original and reflected sound waves. For example, the reflection of sound back up a tube from an open or closed end may result in the reflected wave meeting the remainder of the original wave. The interaction results in *superposition* of the waves, which creates the characteristic sounds of musical instruments and our voices.

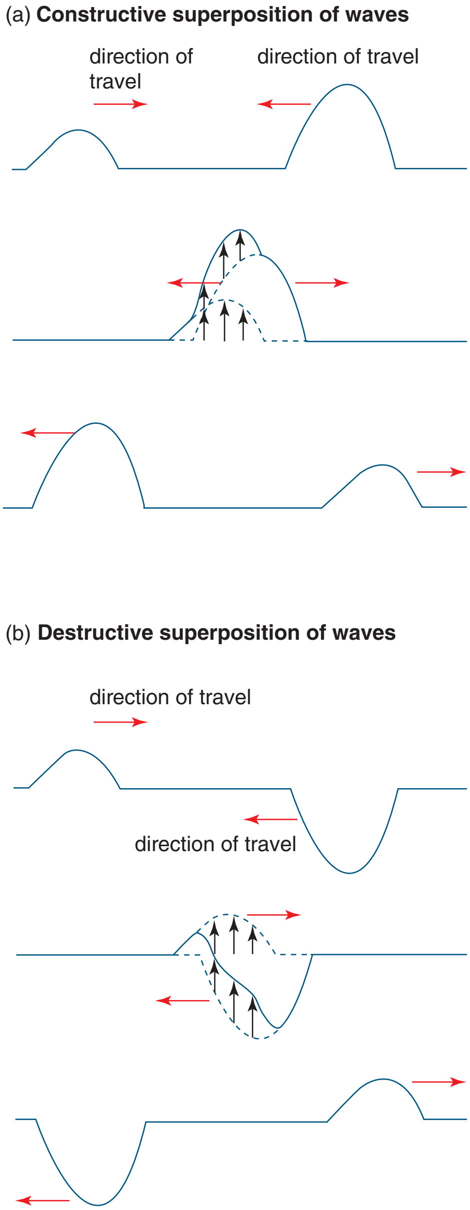


Figure 5.22
Superposition of waves in a string. After the interaction, the pulses continue unaltered: they do not permanently affect each other.

Practical activity
25 Interference of light: Young's double slits

Superposition

Imagine two transverse waves travelling in opposite directions along a string, as shown in Figure 5.22a. When the crest of one wave coincides with the crest of the other, the resulting displacement of the string is the vector sum of the two individual displacements. The amplitude at this point is increased and the shape of the string resembles a combination of the two pulses. After they interact, the two pulses continue unaltered. If a pulse with a positive displacement meets one with a negative displacement, as in Figure 5.22b, a lower amplitude is produced. One wave, in effect, subtracts from the other. Once again, the pulses emerge unaltered.

When two waves meet and combine, there will be places where the resultant displacement of the wave increases and other places where it decreases. The resulting pattern is a consequence of the principle of *superposition*. Although there is a different displacement as the two waves are superimposed, passing through each other does not alter the shape, amplitude or speed of either pulse. Just like transverse waves, longitudinal waves will be superimposed as they interact.

When two or more longitudinal or transverse waves meet, the resulting displacement at each point will be the vector sum of the displacements of the component waves. This is the principle of ***SUPERPOSITION***.

Figure 5.23
The ripples from raindrops striking the surface of a pond behave independently regardless of whether they cross each other. Where the ripples meet, a complex wave will be seen as the result of the superposition of the component waves. After interacting, the component waves continue unaltered.

The effects of superposition are around us all the time. The ripples in the pond of Figure 5.23 were caused by raindrops hitting a pond. Where two ripples meet, a complex wave resulting from the superposition of the two waves occurs, after which the ripples continue unaltered. In a crowded room, all the sounds reaching your ear are superimposed, so that one complex sound wave arrives at the eardrum. Individual sound waves will cross each other repeatedly, but it is still possible to distinguish which person is speaking: you know they will sound the same no matter where you stand. To discern one person's speech amid all the sounds in the room, you use your ability to 'undo' the superposition of waves.

Superposition is important both theoretically and practically in the formation of complex sounds. Imagine two single-frequency sound waves, or pure tones, one of which is twice the frequency of the other. The two individual waves are added together to give a resultant, more complicated, sound wave (Figure 5.24). Where one sound wave has a much greater amplitude, as in the example illustrated, it will still be the predominant sound heard. The quieter, higher frequency sound will combine with the louder one to cause a different quality in the sound that we hear.

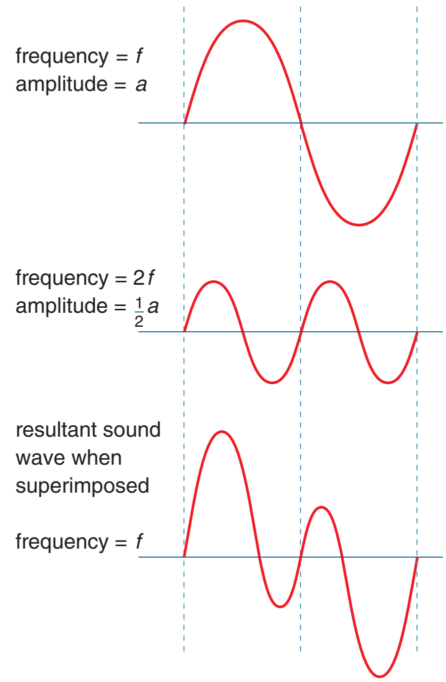


Figure 5.24
Two sound waves, one twice the frequency of the other, produce a complex wave of varying amplitude when they are superimposed.

Physics in action — Some effects of superposition

Superposition creates problems for synthesised sound, but it can also be used to our advantage.

Beats

If two sound waves of equal amplitude but slightly different frequency are combined, the resulting sound has a regular pulsation, called a *beat*. The effect is only noticeable when the two frequencies are close to each other. This phenomenon is the direct result of superposition of the sound waves.

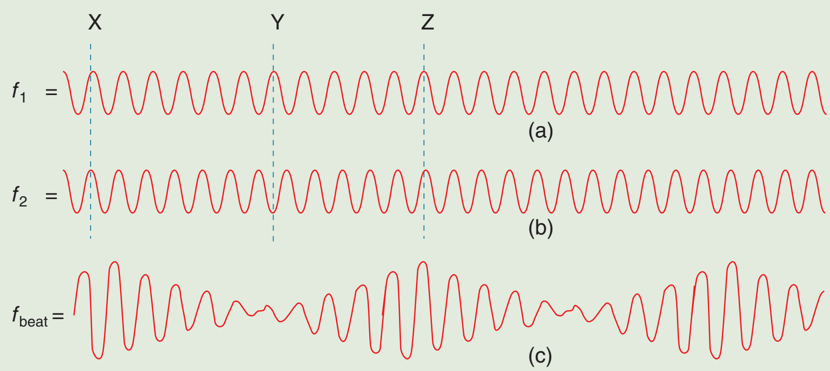


Figure 5.25
A beat (f_{beat}) produced by two sound waves f_1 and f_2 of equal amplitude but slightly different frequency. At time X the waves are in phase, producing a larger amplitude. A short time later, at Y, they are out of phase, resulting in a smaller amplitude, and at Z they are in phase once again. When the two waves are in phase (at X and Z), the resulting pressure variation, and hence volume, is large. At Y the waves are out of phase and the resulting amplitude and volume are zero.

Beats can be explained by looking at the graphs of pressure variation with time for two waves of the same amplitude and a slightly different frequency over time (Figure 5.25).

The frequency of the beats (the *beat frequency*) is simply the difference between the frequencies of the sounds that are superimposed to cause the beats:

$$f_{beat} = |f_1 - f_2|$$

Beats are a reliable means of detecting small differences in the frequency of two sounds, and are therefore useful in tuning musical instruments. While the exact frequency of a note may be hard to determine, notes that are not exactly the same will produce a beat. The instruments are tuned so that the beat gradually slows down, indicating that the two instruments are getting closer to the same frequency. When one sound of constant volume is heard, the instruments are at the same frequency. (Of course, this method does not guarantee that the frequency will be the right one unless a standard tone, electrical tuner or tuning fork is used as a control.)

Beats are also the cause of the oscillating drone of a twin-engined aircraft, produced by the very slightly different speeds of the two propellers.

Synthesisers

An electronic synthesiser attempts to reproduce the sound of musical instruments, voices or natural sounds by combining—or superimposing—many waves of different frequencies and amplitudes. The output signal consists of a number of electrical oscillations that have been mixed in a controlled manner to create a wave which duplicates the original sound as closely as possible. The accuracy of the output depends on how closely the input frequencies match those of the original.

The total range of frequencies produced by a particular musical instrument is referred to as its sound spectrum. Even small factors influencing the construction of the instrument can affect the relative amplitudes of particular frequencies. An analysis of the superimposed waveform needs to consider the amplitude of each frequency and the overall wave envelope (the length of time and particular point in time each frequency is heard). The complex waveforms of three common instruments are shown in Figure 5.26.

Synthesisers mimic wind instruments better than string instruments, because generally there are fewer frequencies to combine. The final sound from a synthesiser also depends on playing style. For example, a piano note has a strong beginning—or attack—and then decays, whereas a violin or flute can be sustained for some time. The many shortcomings of synthesised sounds have led to the development of 'samplers'. Samplers allow digitised recordings of actual instruments to be 'played' instead of synthesised sounds.

Noise reduction

A relatively recent application of the superposition of sound waves has been to combat some of the adverse effects of noisy working environments. Carefully placed microphones in a noisy workplace pick up and record repetitive loud noise. A computer and amplifier reverse the phase of the sound, creating 'antisound'. When the antisound is played through a worker's headphones, it can cancel out the noise by superposition of the sound waves. The volume of the resulting displacement, as heard by the worker, is close to zero.

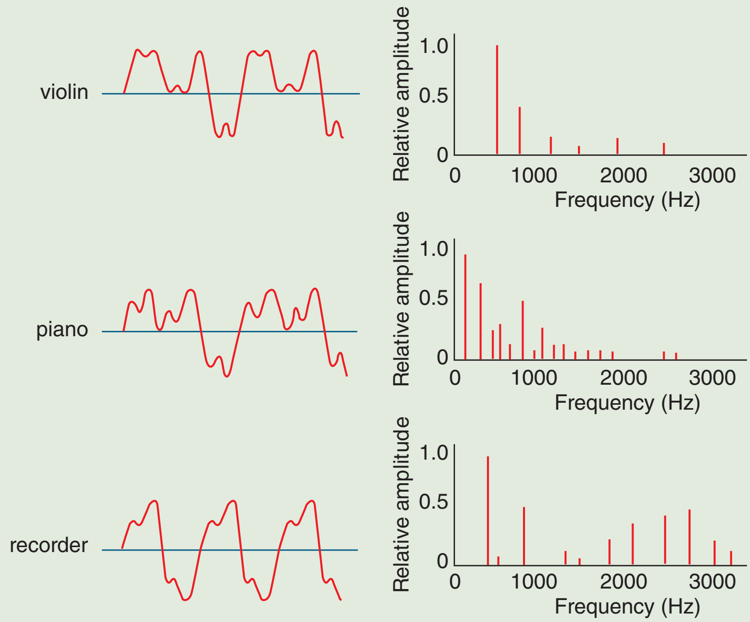


Figure 5.26

The complex waveforms produced by the violin, piano and recorder result from the superposition of a range of frequencies with differing energies and durations.

Resonance

You have probably heard about singers who can break glass by singing particularly high notes. Figure 5.27 shows a glass being broken in much the same way. A loudspeaker is emitting a particular frequency that suddenly destroys the glass. Sounds of a different frequency may have no effect at all, even if they are louder. Any object or system which can vibrate can be made to *resonate* by waves or pulses of exactly the right frequency, called the *resonant frequency*. If the amplitude of the vibrations becomes too great, the object can be destroyed.

Figure 5.27

A glass can be destroyed by the vibrations caused by a speaker emitting a sound of the same frequency as the resonant frequency of the glass.

A swing pushed once and left to swing (*oscillate*) freely is an example of an object oscillating at its *natural frequency*. The frequency at which it moves backwards and forwards depends entirely on the design of the swing—mostly on how long its supporting ropes are. In time, the oscillations will fade away as the energy is transferred to the supporting frame and the air. But if you were to push the swing repeatedly, at a rate chosen by you, each successive push would give more energy to the system—as long as you maintain the same rate. (Pushing out of phase or against the rate you have established would *decrease* the total energy.) This is an example of a *forced vibration*. Regardless of the swing's natural pattern of movement, it is being forced to move at the rate you determine. The swing will continue to move as long as you continue to supply energy.

Resonance occurs when a *forcing frequency* equals the natural frequency of the object in question. If you were to watch the swing to determine its natural frequency of vibration and then push with the same rhythm, the forcing frequency would match its natural frequency. In this case, the additional energy you add by pushing will increase the amplitude of the swing rather than work against it. Over time, the amplitude will increase and the swing will go higher and higher: this is resonance. The swing can only be pushed at one particular rate to get the desired increase in amplitude (i.e. to get the swing to resonate). If the rate is faster or slower, the forcing frequency that you are providing will not match the natural frequency of the swing and you will be fighting against the swing rather than assisting it.

RESONANCE in an object occurs when the forcing frequency equals the natural frequency of the object.
Other examples of resonant frequency that you may have encountered relate to musical instruments: blowing air across the mouthpiece of a flute or bowing a violin in just the right place. In each case, a clearly amplified sound will be heard when the frequency of the forcing vibration matches a natural resonant frequency of the instrument.

Two significant effects occur when the natural resonant frequency of an object is matched by the forcing frequency.

- The amplitude of the oscillations within the resonating object will increase dramatically.
- The maximum possible energy from the source creating the forced vibration is transferred to the resonating object.

You can feel both these effects when you drag a wet finger around the top of a crystal glass. Your finger sets up vibrations in the glass, which you can easily feel. When the frequency of vibration from the finger exactly matches the natural resonant frequency of the glass, you can hear an increase in the intensity of the sound produced. The initially small vibrations are significantly amplified and are easily heard, even across a crowded and noisy room. This example is typical of the way many resonant systems amplify sound.

In musical instruments and loudspeakers, resonance is a desired effect. The sounding boards of pianos and the enclosures of loudspeakers are designed to enhance and amplify particular frequencies. In other systems, such as car exhaust systems, resonance is not always desirable, and care is taken to design a system which prevents resonance.

Physics file

Resonance was responsible for destroying a suspension bridge over the Tacoma Narrows Gorge in the US state of Washington in 1940. Wind gusts of 70 km h^{-1} provided a forcing frequency of vibration which caused the bridge to oscillate with an ever-increasing amplitude, until the whole bridge shook itself apart. Nowadays, bridges and buildings are subjected to wind tunnel tests at the design stage to identify any problems that might occur due to resonance.

Physics file

When a string in a piano is struck, not only will that string vibrate—any string which corresponds to a harmonic of the original (i.e. the same frequency or an exact number of octaves higher) will also vibrate as a result of resonance and add its own sound to the total sound being heard. The principle can be demonstrated by mounting two tuning forks of the same frequency on a sounding box, as shown in Figure 5.28. When one fork is struck, the second fork will also vibrate. The forcing vibration of the sound wave created by the original causes the second to vibrate through resonance.

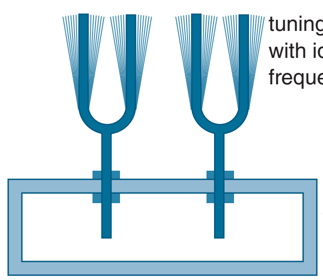


Figure 5.28
When one tuning fork is struck, the second identical fork will vibrate as a result of resonance.

Standing waves

Drawing a bow across a violin string causes the string to vibrate between the fixed bridge of the violin and the finger of the violinist. The simplest vibration will have maximum amplitude at the centre of the string, halfway between bridge and finger. This is a very simple example of a *transverse standing wave*. At both the bridge and finger, the amplitude of vibration will always be zero (a node), since the string is fixed at these points. Halfway between the two nodes, the amplitude of the wave will be maximum (an antinode). In a standing wave, the nodes and antinodes remain stationary (hence the name stationary or *standing wave*). This situation contrasts with a *travelling wave*, where every point on the wave has a maximum displacement at some time. The simplest form of standing wave, or *mode of vibration*, is shown in Figure 5.30. It has only two nodes, and so the length of the string, L, will correspond to $\frac{\lambda}{2}$.

Figure 5.29
The sound box of an acoustic musical instrument is tuned to resonate for the range of frequencies of the forcing vibrations being produced by the strings. When a string is plucked or bowed, the airspace inside the box vibrates in resonance with the natural frequency and the sound is amplified. Loudspeaker enclosures are designed on the same principle.

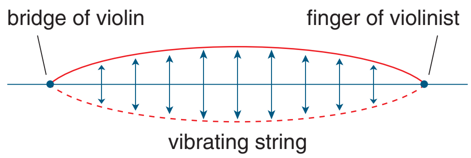


Figure 5.30
The simplest mode of vibration of a bowed violin string is an example of a transverse standing wave.

Standing waves are an important phenomenon of the superposition of waves. They occur when two waves of the same amplitude and frequency are travelling in opposite directions in the same string. Usually, one wave is the reflection of the other. Standing waves are responsible for the wide variety of sounds we associate with speech and music. They occur in all stringed musical instruments. As a string is plucked, struck or bowed, a great variety of vibrations are created. This complex set of vibrations can occur simultaneously. Each wave will propagate through the length of the string and reflect from the fixed ends. Most of the reflections will interfere in a purely random manner and will quickly die away. Particular frequencies will reflect in such a way that standing waves are created, which resonate and combine to give a particular instrument its characteristic sound. Figure 5.32 shows how this can occur. The solid line represents the original transverse wave train within a string, while the broken line represents its reflection from an end. Both wave trains are of the same amplitude and frequency, travelling in opposite directions with the same speed.

Physics file

It is important to realise that the term 'standing wave' does not mean that the string is stationary. In fact, it continues to oscillate; it is the relative position of the nodes and antinodes that remains unchanged. If the various stages during the oscillation of a particular standing wave are superimposed, the sequence of movement may look something like Figure 5.31. It is also important to note that standing waves are *not* a natural consequence of every wave reflection:

- They can only be produced by the superposition of two waves of equal amplitude and frequency, travelling in opposite directions.
- They are the result of resonance and occur only at the natural frequencies of vibration of the string.

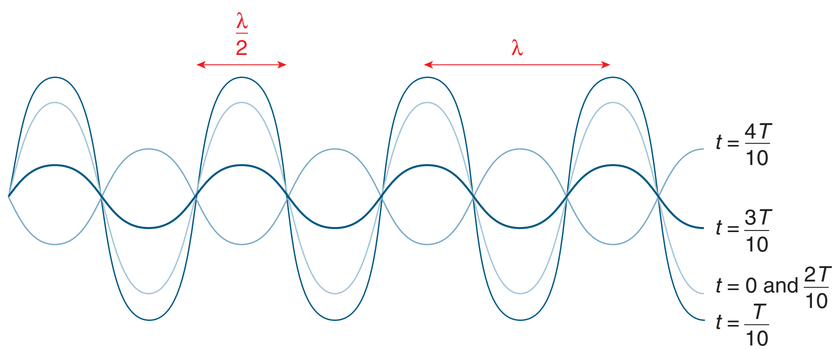


Figure 5.31
Within the wave envelope, a transverse standing wave may appear to have a variety of amplitudes, from $-2A$ to $+2A$. However, the nodes remain stationary.

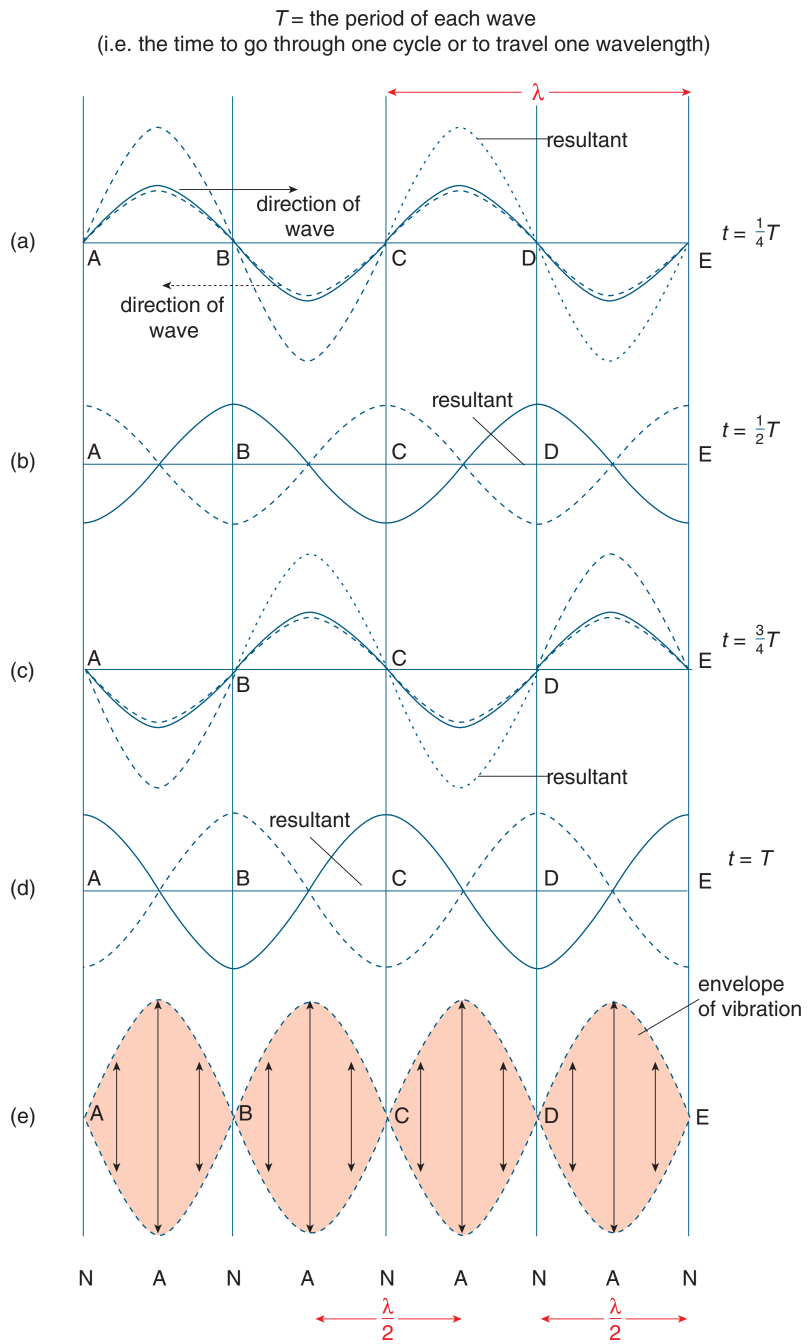


Figure 5.32

A standing (stationary) wave created in a string from two waves travelling in opposite directions, each with the same amplitude and frequency.

(a) At a particular point in time, the two waves are completely superimposed so that crest meets crest and trough meets trough, to give the wave twice the original amplitude.

(b) After a time equal to $T/4$ (one-quarter of the period) the waves will have moved $\lambda/4$, which means that they have moved $\lambda/2$ in relation to each other. The waves are completely out of phase and the resulting displacement is zero.

(c) and (d) As further time goes by, the waves will continue to move past each other, creating the superimposed waveforms.

(e) In the final wave envelope, the standing wave swings between maximum displacements, creating antinodes (A) which lie halfway between the stationary nodes (N). Regardless of the position of the component waves, these nodes stay in the same place. Successive nodal points lie $\lambda/2$ apart, as do successive antinodal points.

Harmonics

The frequencies produced in the complex vibration of standing waves in a stringed instrument are called *harmonics*. The simplest mode of vibration, which has only one antinode, is called the *fundamental*. Higher-level modes of vibration are referred to as harmonics *or overtones*. Figure 5.33 illustrates the first few harmonics in a string of fixed length which is plucked, struck or bowed. The fundamental frequency usually has the greatest amplitude, so it has the greatest influence on the sound of the note. The amplitude generally decreases for each subsequent harmonic. All harmonics are usually produced in a string simultaneously,

Physics file

Because of the relationship between the length of a string and the wavelength of the sound produced, it is easy to see the importance of finger placement when playing a string instrument. What is not so apparent is the importance of the position at which the string is plucked or bowed.
The seventh harmonic of a violin string produces the high-pitched squeal commonly made by beginners. The solution is to bow the strings at the position of the first node for the seventh harmonic: then the standing wave of this harmonic cannot form and the squeal is not heard. The standing waves corresponding to the lower harmonics require longer distances to the first nodal position and will not be affected. (Try calculating the distance to the first nodal point of the seventh harmonic on a real violin, and test this theory.)

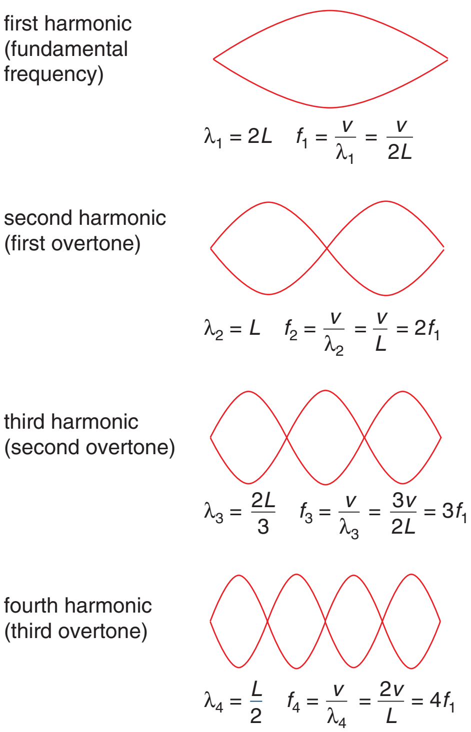


Figure 5.33
The first four possible harmonics in a stretched string. The fundamental, or first harmonic, usually has the largest amplitude. The ends are fixed, so they will always be nodal points.

and the instrument and the air around it also vibrate to create the complex mixture of frequencies we hear as an instrumental note.

The harmonics represent the resonant frequencies for the string. They can be calculated from the relationship between the length of the string L and the wavelength of the corresponding standing wave. Since the first harmonic matches the length of the string, $L = \frac{\lambda}{2}$. The second and third harmonics satisfy the conditions $L = \frac{2\lambda}{2}$ and $L = \frac{3\lambda}{2}$ respectively.

Wavelength of the harmonics in a ***STRING***:

$$L = \frac{n}{2} \times \lambda_n$$

$$\lambda_n = \frac{2L}{n}$$

where λ_n is the wavelength of the nth harmonic (m), L is the length of the string (m) and n is the number of the harmonic = 1, 2, 3, . . .

Using the wave equation, $v = f\lambda$, the relationship between frequency, velocity and string length can be established:

Frequency of the harmonics in a ***STRING***:

$$f_n = \frac{nv}{2L}$$

where f_n is the frequency of the nth harmonic (Hz) and v is the velocity of the component waves (m s^{-1}).

✓ Worked Example 5.3A

When the key on a piano corresponding to a note of 440 Hz is struck firmly, it is found that the 880 Hz string also vibrates, even though the two are not connected and are some distance apart. How can this be so?

Solution

This situation involves resonance. The frequency of the 880 Hz string corresponds to the second harmonic of the forcing vibration by the 440 Hz string (i.e. it is double the frequency). Hence, it will resonate at its natural frequency of vibration. The sound produced by the 440 Hz string provides the necessary energy.

✓ Worked Example 5.3B

A violin string has a length of 22.0 cm. It is vibrating with its fundamental mode of vibration at a frequency of 880 Hz. What is the wavelength of this fundamental frequency?

Solution

The fundamental frequency corresponds to the first harmonic. That is:

$n = 1 \qquad \lambda_n = \frac{2L}{n}$

$L = 22.0 \times 10^{-2}$ m $\qquad \lambda_1 = \frac{2(22.0 \times 10^{-2})}{1}$

$= 4.40 \times 10^{-1}$ m

Wind instruments and air columns

Longitudinal stationary waves are also possible in air columns. These create the sounds we associate with wind instruments. Blowing over the hole of a flute or the reed of a saxophone produces vibrations that correspond to a range of frequencies that create sound waves in the tube.

The compressions and rarefactions of the sound waves, confined within the tube, reflect from both open and closed ends. This creates the right conditions for resonance and the formation of standing waves. The length of the pipe will determine the frequency of the sounds that will resonate.

Open-ended air columns

At the open end of a pipe, sound waves are reflected. When a compression or rarefaction reflects from an open end, it does so with a phase change of $\lambda/2$. This causes the reflected wave to destructively interfere with the incoming wave, and a pressure node results. This point is at air pressure.

In a pipe that is open at both ends, a standing sound wave can be produced with a pressure node at each end. This means that the harmonics of an open-ended pipe will be very similar in nature to those of a string fixed at both ends. The wavelength of the fundamental frequency (the first harmonic) will be twice the effective length of the air column. That is:

Wavelength of the harmonics in a ***PIPE OPEN AT BOTH ENDS***:

$$\lambda_n = \frac{2L}{n}$$

where λ_n is the wavelength of the nth harmonic (m), L is the effective length of the air column (m) and n is the number of the harmonic = 1, 2, 3, . . .

Using the wave equation, $v = f\lambda$, the relationship between frequency, velocity and air column length can also be established:

Frequency of the harmonics in a ***PIPE OPEN AT BOTH ENDS***:

$$f_n = \frac{nv}{2L}$$

where f_n is the frequency of the nth harmonic (Hz) and v is the velocity of the component waves (m s^{-1}).

All harmonics are possible in an air column that is open at both ends. The first few harmonics are illustrated in Figure 5.34. Just as for strings, the frequency of the harmonics will be whole multiples of the fundamental ($f_1 : f_2 : f_3 = 1 : 2 : 3$). The second harmonic has twice the frequency of the fundamental ($f_2 = 2f_1$), the third has three times the frequency ($f_3 = 3f_1 = \frac{3}{2}f_2$), and so on.

Typical open-ended pipes include flutes, oboes and similar instruments and the muffler in a car exhaust system.

Physics file

Wind instruments do not have the benefit of having strings of different masses that can be tensioned or released to produce different frequencies. In bugles or ceremonial trumpets, where the length is fixed, the pitch of the note produced must come from the range of harmonics available as a direct result of the length of the tube. This severely restricts the number of notes that can be played. Other instruments allow the player to vary the length of the pipe by covering or uncovering holes along the length of the pipe or, in the case of orchestral trumpets, using valves to connect additional curved lengths of pipe. Coiled lengths of pipe in trombones, French horns and trumpets allow a greater total length and thus a greater potential range of notes. Part of the skill of a woodwind or brass musician is controlling which mode of vibration dominates the final sound.

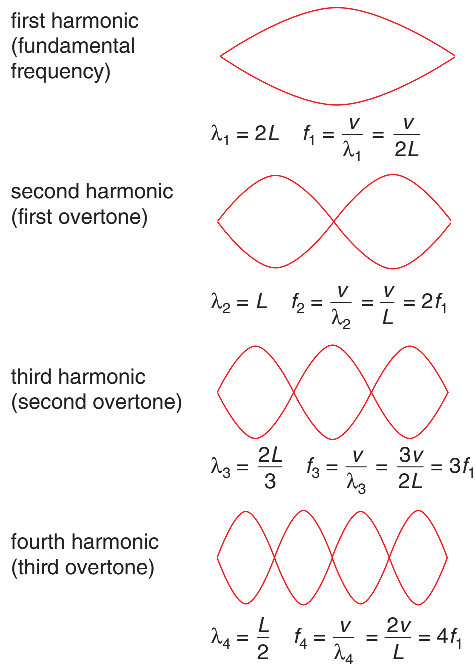


Figure 5.34
The first harmonics for a pipe that is open at both ends. The two lines represent the envelope of the standing wave. Note that the effective length of the air column carrying the sound is longer than the actual pipe length. For a pipe open at both ends all harmonics are possible, and the ratio $f_1 : f_2 : f_3$ is 1 : 2 : 3.

Figure 5.35

The flute is a typical example of a pipe open at both ends where an air column can be made to vibrate.

Practical activity

19 Visualising standing waves in air columns

Physics file

If the end of a pipe is flared (e.g. in a trumpet or saxophone) the situation is far more complex than for a plain end. As a wavefront leaves the relatively constant diameter of the main tube and enters the horn region of the pipe, it begins to spread and behave in a different manner. This maximises the amount of energy transmitted to the surrounding air. Research is still continuing on exactly how sound behaves in this region.

Practical activity

20 Speed of sound by resonance tube

✓ Worked Example 5.3C

A particular flute has an effective length of 35.0 cm. It can be thought of as an open-ended air column. The sound in the tube has a velocity of 350 m s^{-1}.

a What is the wavelength of the second harmonic?

b What will the frequency of the second harmonic be?

Solution

a For the second harmonic, a drawing of the appropriate standing wave will show that:

$L = 35.0 \times 10^{-2}$ m

$n = 2$

$$\lambda_n = \frac{2L}{n}$$

$$\lambda_2 = \frac{2(35.0 \times 10^{-2})}{2}$$

$$= 3.50 \times 10^{-1} \text{ m}$$

b Using the wave equation and rearranging:

$n = 2$

$L = 35.0 \times 10^{-2}$ m

$v = 350$ m s^{-1}

$$f_n = \frac{nv}{2L}$$

$$f_2 = \frac{2(350)}{2(35.0 \times 10^{-2})}$$

$$= 1.00 \times 10^{3} \text{ Hz}$$

Closed air columns

In this section, a 'closed air column' means that the pipe or tube is closed only at one end and remains open at the other. This situation is different from that in strings and fully open pipes; in both those cases, the reflection of the wave is the same at both ends. The open end of the air column reflects a sound wave with a change of phase, creating a pressure node. However, at the closed end, there will be no change of phase for a reflected sound wave, so here the reflected waves interfere constructively with the incoming waves and a pressure antinode occurs. Air particle movement will be minimal in this region.

So standing waves established in a closed tube will have a node at one end and an antinode at the other (see Figure 5.36).

The simplest harmonic, or fundamental frequency, will have a wavelength four times the length of the effective air column. The next simplest harmonic will have a wavelength $\frac{4}{3} \times L$ of the air column; the next, $\frac{4}{5} \times L$; and so on. In general:

Wavelength and frequency of the harmonics in a ***PIPE CLOSED AT ONE END***:

$$\lambda_n = \frac{4L}{n}$$

and

$$f_n = \frac{nv}{4L}$$

where λ_n is the wavelength of the harmonic (m), L is the effective length of the air column (m), f_n is the frequency of the nth harmonic (Hz), v is the velocity of the component waves (m s^{-1}) and n is the number of the harmonic = 1, 3, 5, . . . (odd numbers only).

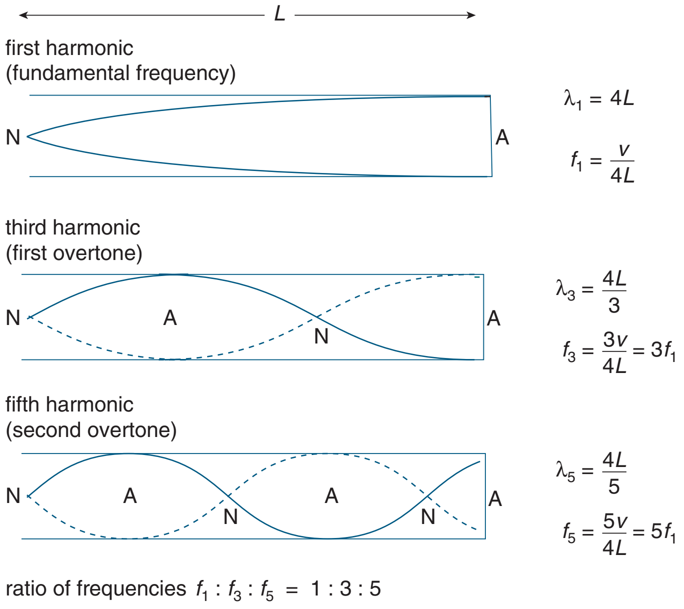


Figure 5.36

The lower harmonics for a pipe that is closed at one end. The two lines represent the maximum pressure variation of the standing wave. Only odd-numbered harmonics are possible, since only these satisfy the condition of having a pressure node at the open end and a pressure antinode at the closed end.

Notice that in this situation only odd-numbered harmonics are possible. Figure 5.36 illustrates the standing waves in terms of pressure variation for a closed tube. The ratio of one frequency to another is $f_1 : f_3 : f_5 = 1 : 3 : 5$. That is, the third harmonic is three times the frequency of the fundamental ($f_3 = 3f_1$), the fifth harmonic is five times the frequency ($f_5 = 5f_1 = \frac{5}{3}f_3$), and so on.

There is no second or fourth harmonic.

Some examples of closed air columns are the human vocal tract, the ear canal, ported loudspeakers, car engine manifolds and some organ pipes.

✓ Worked Example 5.3D

The ear canal, from the outer ear to the eardrum, can be thought of as a tube closed at one end (by the eardrum) and open at the other. It is approximately 3.00 cm long in an adult. Assume that the speed of sound is 340 m s^{-1}.

a What is the fundamental resonant frequency of the ear canal?
b What is the frequency of the next resonating frequency?
c Explain one reason why some frequencies are heard better than others.

Solution

a It is useful to refer to a diagram of the situation for a particular harmonic in order to determine the relationship between λ and L. Figure 5.36 shows that for a closed tube, the length will correspond to $L = \frac{\lambda}{4}$ and so:

$v = 340$ m s^{-1} $\qquad f_n = \frac{nv}{4L}$

$L = 3.00 \times 10^{-2}$ m $\qquad f_1 = \frac{1(340)}{4(3.00 \times 10^{-2})}$

$n = 1$ $\qquad = 2.83 \times 10^3$ Hz

Compare this with the frequencies heard best by humans.

b As this is a closed tube, only odd-numbered harmonics will form, so the next harmonic will be the third:

$v = 340 \text{ m s}^{-1}$

$L = 3.00 \times 10^{-2} \text{ m}$

$n = 3$

$$f_n = \frac{nv}{4L}$$

$$f_3 = \frac{3(340)}{4(3.00 \times 10^{-2})}$$

$$= 8.50 \times 10^3 \text{ Hz}$$

c The standing waves of resonant frequencies will have substantially larger amplitudes than the sounds that do not cause resonance. As a result, sounds which closely correspond to resonant frequencies, like those calculated above, will be amplified more and have more chance of being heard than others. Of course there are many other factors which influence the range of frequencies a particular person is able to hear.

Physics in action — Music or noise?

What is music to one person may be just plain noise to another. The frequencies that are present in sounds vary, and any clear distinction between music and noise is mainly a subjective judgement. Physics can provide a basic distinction between what we define as a harmonious sound and as noise, but it does not attempt to draw a line between the two.

Related and unrelated frequencies

Musical instruments generally produce sounds which are whole-number multiples of a fundamental frequency. When a sound from a musical instrument is displayed on an oscilloscope or a digitiser, a clear and distinctive stable wave pattern is seen. This is because the vibrations produced by most musical instruments are only free to vibrate in one dimension. A sound which includes these simple multiples of the fundamental frequency is usually regarded as harmonious.

Hitting two sticks together creates a different mixture of frequencies. The vibrations within the sticks that cause the sound move in three dimensions, and the range of frequencies produced is more complex. The display on a digitiser or an oscilloscope would not show a stable wave pattern; the frequencies bear little or no relation to each other. The sound may still be quite pleasant to hear, but it can be interpreted as noise since it consists of a random mixture of frequencies. 'White noise' is such a mixture of frequencies. Although we define it as noise, it can be soothing and is quite a pleasant sound to many people.

Instruments such as drums lie between the extremes of simple multiples of frequencies and random mixtures. Drums can vibrate in two dimensions, so they do not have the pure harmonious sound of other instruments, but the frequencies do bear some relation to each other. Totally enclosed drums can still sound harmonious, because certain frequencies resonate and are accentuated.

Harmonies and musical scales

Harmonies are an important part of music. Two notes generally sound harmonious if the ratio of their frequencies is a simple whole number. The simplest ratio of 2:1 sounds the most harmonious: this is what we call an octave. A ratio of 3:2, called a 'fifth' since it includes five musical intervals (tones and semitones), also sounds very harmonious. Combinations based on more complex ratios, such as 5:4, sound 'dissonant'. Modern music has seen the acceptance of more dissonant combinations, and today almost any dissonance can be found in a musical score. Many different musical scales involving these ratios have been used throughout the world. For example, African music uses much smaller intervals between notes and thus has two to four times as many defined notes as traditional Western music.

Figure 5.37
Because the vibration of a drum skin occurs in two dimensions, in contrast to the single longitudinal dimension of a string, it creates less harmonious sounds than many other musical instruments.

The most common modern scale is called the 'equally tempered chromatic scale'. This evolved from a scale based on four intervals corresponding to simple fractions of the length of a vibrating string. Various versions of the earliest chromatic scale led to a compromise that suits keyboard instruments such as the piano. (Earlier versions had many more intervals within one octave and would have required a large and unplayable keyboard.) The equally tempered chromatic scale divides a simple 2:1 ratio, one octave, into 12 equal semitones. Each pair of adjacent notes thus has the same frequency ratio.

As an octave is a basic ratio of 2:1, having 12 equal intervals means that the ratio of the frequency of any one note on a piano to that of the next will be $\sqrt[12]{2}:1$ or about 1.06:1.

table 5.1 Frequencies of common notes on the equally tempered chromatic scale, based on A = 440 Hz The tonal names refer to the seven white notes on the keyboard. Compromises in the scale mean that adjacent sharps (#) and flats (*b*) have the same frequency. (Upper C actually has exactly double the frequency of middle C. It is noted here as 523 and not 524 Hz because the values are rounded to the nearest whole number.)

Note	Tonal name	Frequency (Hz)
A		220
$A^{\#}$, B^{b}		233
B		247
C (middle)	do	262
$C^{\#}$, D^{b}		277
D	re	294
$D^{\#}$, E^{b}		311
E	mi	330
F	fa	349
$F^{\#}$, G^{b}		370
G	so	392
$G^{\#}$, A^{b}		415
A	la	440
$A^{\#}$, B^{b}		466
B	ti	494
C	do	523

5.3 SUMMARY Wave interactions

- The principle of superposition tells us that when two or more waves interact, the resultant displacement or pressure at each point along the wave will be the vector sum of the displacements or pressures of the component waves.
- Resonance occurs when the frequency of a forcing vibration equals the natural frequency of an object.
- Two special effects occur with resonance: (i) the amplitude of vibration increases, and (ii) the maximum possible energy from the source is transferred to the resonating object.
- Standing, or stationary, waves occur as a result of resonance at the natural frequency of vibration.
- Standing waves are the result of superposition of two waves of equal amplitude and frequency, travelling in opposite directions in the same medium.
- The standing wave frequencies are referred to as harmonics. The simplest mode is referred to as the fundamental frequency.
- Within a string, the wavelength of the standing waves corresponding to the various harmonics is $\lambda_n = \frac{2L}{n}$, with $f_n = \frac{nv}{2L}$. All harmonics ($n = 1, 2, 3, \ldots$) may be present, and the ratio of frequencies is $f_1 : f_2 : f_3 : \ldots = 1 : 2 : 3 : \ldots$

- In a tube that is open at both ends, the wavelength of the fundamental frequency is twice the effective length of the air column, i.e. $\lambda_n = \frac{2L}{n}$, and $f_n = \frac{nv}{2L}$ where n is the number of the harmonic = 1, 2, 3, . . . All harmonics are possible, and the ratio of frequencies is $f_1 : f_2 : f_3 : \ldots = 1 : 2 : 3 : \ldots$
- In a tube with one closed end, the fundamental frequency will have a wavelength four times the length of the effective air column. In general: $\lambda_n = \frac{4L}{n}$, with $f_n = \frac{nv}{4L}$ where n is the number of the harmonic = 1, 3, 5, . . . Only odd-numbered harmonics are possible, and the ratio of frequencies is $f_1 : f_3 : f_5 : \ldots = 1 : 3 : 5 : \ldots$
- A rise in tone of one octave is equivalent to a doubling of the frequency.

5.3 Questions

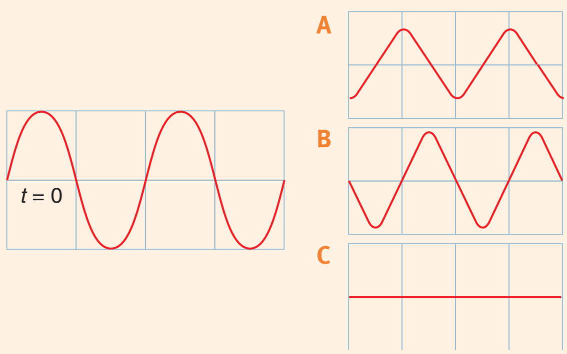
1 State if each of the following statements concerning the interaction of two pulses is true or false.
  a The displacement of the resultant pulse is equal to the sum of the displacements of the individual pulses.
  b As the pulses pass through each other, the interaction permanently alters the characteristics of each pulse.
  c After the pulses have passed through each other, they will have the same characteristics as before the interaction.
  d The interaction will cause a loss in kinetic energy of each pulse, resulting in the pulses separating with reduced velocities.

2 Why it is possible to shatter a glass by exposing it to sound of a certain frequency?

3 How does an acoustic guitar amplify the sound it produces?

4 Explain how a standing wave is produced.

5 The following diagram shows the stationary wave pattern produced in a vibrating string at time $t = 0$. The period of the stationary wave is T.



Which of the diagrams A–C best describes the shape of the standing wave at the following times?
  a $t = \frac{T}{4}$
  b $t = \frac{T}{2}$
  c $t = \frac{3T}{4}$

6 At what point should a guitar string be plucked to make its:
  a fundamental frequency most prominent?
  b second harmonic most prominent?
  c third harmonic most prominent?

7 The speed of a transverse wave in a metal string of 50.0 cm length, when subjected to a certain tension, was calculated to be 300.0 m s^{-1}. If this string were to be plucked, calculate the frequency of the:
  a first harmonic
  b second harmonic
  c third harmonic.

8 Very briefly explain how resonance occurs in an air column.

9 A flute can be considered to be an open-ended air column. Consider a flute of effective length of 45.0 cm.
  a What is the wavelength of the fundamental note produced by the flute?
  b What is the wavelength of the second harmonic produced by the flute?
  c Calculate the frequency of the third harmonic produced by the flute. The speed of sound inside the flute is 330.0 m s^{-1}.

10 An organ pipe is an air column closed at one end with an effective length of 75.0 cm. The speed of sound inside the pipe is 330.0 m s^{-1}.
  a What is the frequency of the fundamental note produced by the pipe?
  b What is the frequency of the third harmonic?
  c What are the frequencies of the next two harmonics (after the third) that the pipe can produce?

5.4 Electromagnetic radiation

The beginning of a model for light

Now that some of the details of wave motion are once again familiar, we can begin to discuss whether light can be considered a wave, or whether we can use ray optics and particle ideas to model everything that light does!

A **model** is a system of some type that is well understood and that is used to build a mental picture or analogy for an observed phenomenon, in our case the behaviour of light. A good model will appear to behave in the same manner as the entity being investigated. A model needs to be able to explain the observations of light that have already been made and ideally it would predict new behaviours for light. Therefore when deciding upon a model for light we must first examine what we know about it already.

The beginning of human interest in the nature of light dates back to the ancient Greek, Arabian and Chinese philosophers. These early thinkers provided us with the foundations for many of our ideas about mathematics, science, philosophy, architecture, politics and literature over 2000 years ago. They knew that light travels in straight lines and they used this principle widely in surveying and astronomy. The idea was certainly promoted by the mathematician Euclid (*c.* 280 BCE), who included it in his book *Optica*. Furthermore, the law of reflection and an approximation to the law of refraction were both taught by Ptolemy around 150 CE. The Greeks understood that these laws applied both on the Earth and in the cosmos generally, and today this remains an important assumption.

Therefore, during the 17th century, when Newton was studying the nature of light, it was known that light:

- travels in straight lines in a uniform medium (linear propagation)
- obeys the laws of reflection
- obeys Snell's law of refraction.

Electromagnetic waves

What *is* light? In the late 1600s it was known to involve the transfer of energy from one place to another. In Isaac Newton's time a corpuscular (particle) model and a wave model for light had seemed equally valid. We have discussed these two proposed models of light along with their respective explanations of the *reflection* and *refraction* of light. In spite of considerable endeavour by scientists it was not until the early 1800s that one model prevailed. Thomas Young discovered that sources of light were able to *interfere* with each other just like sound waves and water waves do. This finding led to a universally accepted *wave theory* for light. Furthermore, the speed of light could be measured for the first time, and the wave model of refraction was validated. Meanwhile another area of physics had been developing. By the 1860s investigations being carried out on different forms of **electromagnetic radiation** led to the finding that visible light itself is just one of the many forms of electromagnetic radiation (EMR).

Electricity and magnetism were once considered to be separate subjects. However, moving charges create magnetic fields. Similarly a changing magnetic field can be used to create electricity. In 1864 James Clerk Maxwell used mathematical equations to describe how charges moving periodically in a conductor would set up alternating electric fields and magnetic fields in the nearby region. Maxwell knew that the magnetic and electric fields travelled through space. He calculated their speed and found it to be 300 000 km s^{-1}, exactly the same as the speed of light! Also, he devised mathematical expressions to describe the magnetic and electric fields. The solution to these expressions was found to be the *equation of a wave*. Maxwell had shown that light is an electromagnetic wave.

Today we know that the **electromagnetic spectrum** includes a wide range of frequencies (or wavelengths). All electromagnetic waves are created by accelerating charges which result in a rapidly changing magnetic field and electric field travelling out from the source at the speed of light, as shown in Figure 5.38. Note that the electric field component and the magnetic field component are at right angles to each other and to their direction of travel. Electromagnetic radiation meets the description of a transverse wave.

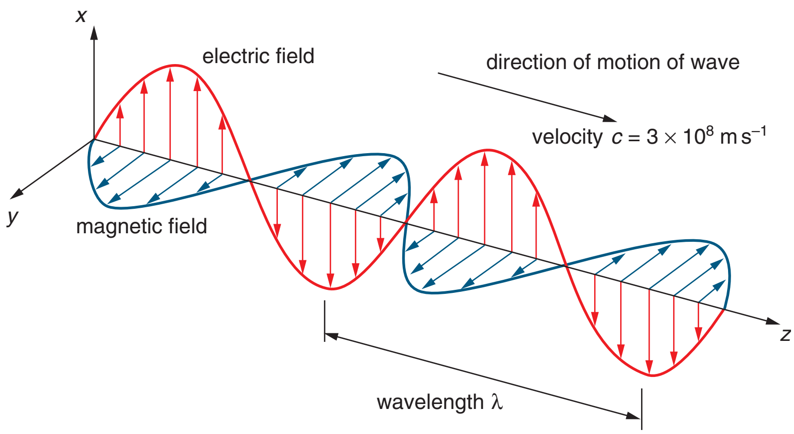


Figure 5.38

Since all electromagnetic waves travel with the same velocity the only thing that differentiates one form of EMR from another is the frequency (and, therefore, the wavelength).

The many forms of EMR are essentially the same, differing only in their frequency and, therefore, their wavelength. The electromagnetic spectrum is roughly divided into seven categories depending on how the radiation is produced and the frequency. The energy carried by the electromagnetic radiation is proportional to the frequency. High-frequency short-wavelength *gamma* rays are at the high-energy end of the spectrum. Low-frequency long-wavelength *radio waves* carry the least energy. Humans have cells in their eyes which can respond to EMR of frequencies between approximately 400 THz and 800 THz (a terahertz is 10^{12} Hz); these frequencies make up the **visible light** section of the electromagnetic spectrum.

Recall that for any wave the relationship between its frequency and its wavelength is given by $v = f\lambda$.

All electromagnetic radiation travels at a speed of 3.00×10^8 m s^{-1} in a vacuum and so this significant speed has been allocated the symbol *c*.

For all ***ELECTROMAGNETIC RADIATION***:

$$f = \frac{c}{\lambda}$$

where f is the frequency of the EMR (Hz), c is the speed of the EMR (3.00×10^8 m s^{-1}) and λ is the wavelength of the EMR (m).

Figure 5.39 shows the different categories of EMR. Note the range of frequencies and wavelengths is enormous. The range of frequencies (or wavelengths) constituting visible light occurs near the middle of the spectrum. Our eyes cannot perceive any wavelengths of EMR outside this range.

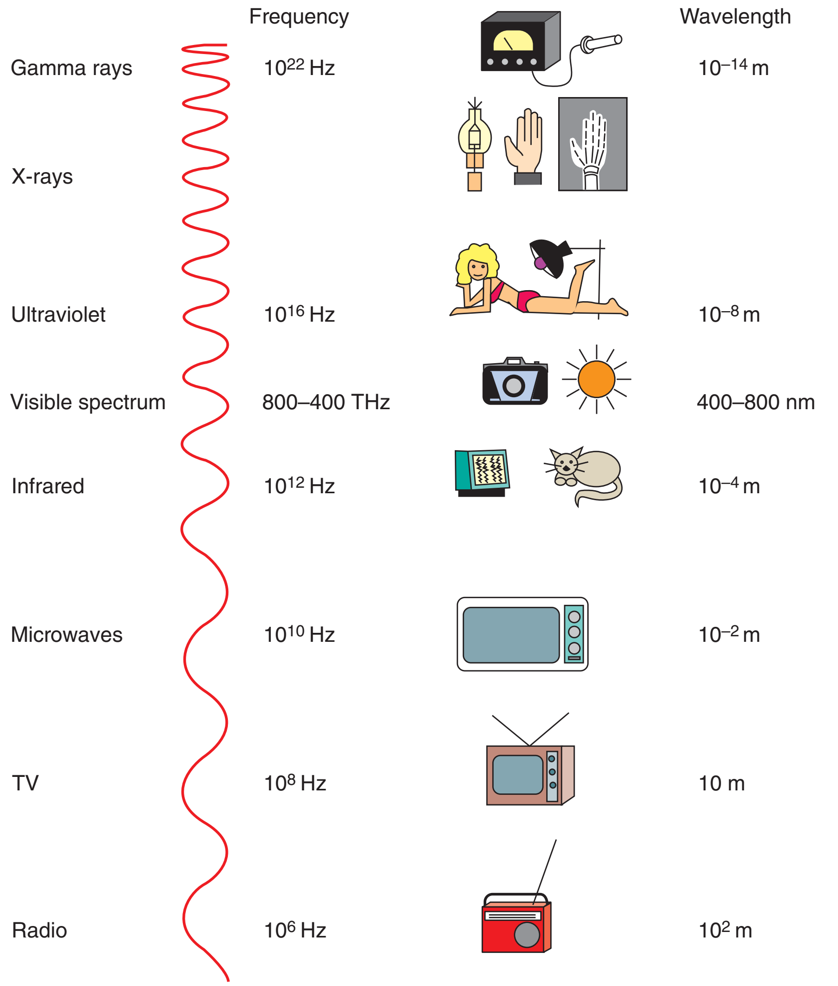


Figure 5.39
The electromagnetic spectrum.

Practical activity
29 Polarisation effects with light

✓ Worked Example 5.4A

The EMR given off by a sample of sodium as it is burned has a wavelength of 589 nm. What is the frequency of this radiation? How would we detect the radiation?

Solution

$\lambda = 589 \times 10^{-9}$ m

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

$$f = \frac{c}{\lambda}$$

$$= \frac{3.00 \times 10^{8}}{589 \times 10^{-9}}$$

$$= 5.09 \times 10^{14} \text{ Hz}$$

This frequency of EMR lies in the visible light section of the electromagnetic spectrum, therefore we would see it! It is actually yellow light.

Physics in action — Other forms of EMR

Radio waves

Accelerating a positive or negative charge can produce EMR. Electrons oscillating in a conducting wire, such as an antenna, produce the radio waves that bring music to your home. The long-wavelength low-energy electromagnetic waves blanket the surrounding region, and aerials can receive the signal many kilometres from the source. As a result of the radio waves, electrons in the receiving aerial wire will oscillate, producing a current that can be amplified. Radio waves can be transmitted over very long distances, including around the Earth's surface, by reflection from layers in the atmosphere.

Microwaves

Microwaves are EMR of wavelengths ranging from about 1 mm to about 10 cm. The microwaves that cook your dinner are produced by the spin of an electron or nucleus. Microwave links are used to allow computer systems to communicate remotely, and radar equipment uses microwave frequencies of centimetre wavelengths.

Infrared waves

Infrared or heat radiation includes the wavelengths that our skin responds to. When you feel the warmth from the Sun or an electric bar heater you are actually detecting infrared radiation. All objects that are not at a temperature of absolute zero radiate EMR. The hotter the object the more radiation is emitted, and the further along the spectrum the radiation is. Night scopes and infrared spy satellites create an image by sensing infrared radiation and converting it into a visible picture.

Ultraviolet waves

Ultraviolet waves have wavelengths shorter than violet light—so our eyes cannot detect them—but no greater than about 10 nm. Many insects can detect the ultraviolet light that is commonly reflected from flowers. Although ultraviolet light is less energetic than gamma rays or X-rays, it is known to cause skin cancer, particularly with increased exposure.

Silicon atoms are able to absorb some frequencies in the ultraviolet region of the spectrum, reducing your chances of getting sunburnt through glass.

X-ray waves

X-rays are produced when fast-moving electrons are fired into an atom. The name is a result of scientists not knowing what they were when they were first detected, hence the letter 'X'. X-rays can pass through body tissue and be detected by photographic film, and so are used in medical diagnosis. They have extensive safety testing, security and quality control applications in industry.

Gamma-ray waves

The highest energy, smallest wavelength radiation is the gamma ray, which is produced within the nucleus of an atom. Gamma rays are one of the three types of emissions that come from radioactive (unstable) atoms. Gamma rays are extremely penetrating and require dense material to absorb them.

Photoelectric effect: Counterevidence for the wave model

Towards the end of the 19th century, physics had entered a most confident era. It was beginning to be believed that it would only be a matter of time before all things could and would be known. During the century following Young, many startling advances had been made, and it was thought that it was simply a matter of continuing in this fashion. But there were a few 'clouds' on the horizon.

One cloud was the **photoelectric effect**. In 1887, when he was experimenting with the production and detection of radio waves, Heinrich Hertz noticed that he could make a spark jump further if the metal surface from which the spark came was illuminated with ultraviolet light. As we will see, a wave interpretation for light would not predict this. So it happened that an experiment challenged an accepted theory; either the theory was wrong or it needed significant modification. This was the seed for the upheaval in physics which ultimately led to the quantum theory. This experimental result, together with Michelson and Morley's celebrated measurement of the speed of light, formed the doorway that led from classical physics to what is referred to as modern physics.

Practical activity

23 Light and a continuous spectrum

The photoelectric effect experiment

The photoelectric effect is the ejection of electrons from the surface of a material when light of a sufficiently high frequency shines upon it. Usually, the electrons are emitted from a metal and they are called '**photoelectrons**'—a term acknowledging light as the reason for their freedom.

An experimental arrangement illustrating all the essential aspects of the photoelectric effect is shown in Figure 5.40. Here, a clean metal surface—the **cathode**—is illuminated with light from an external source. If the light causes photoelectrons to be emitted, they will be detected at the anode. This flow of electrons is called the *photoelectric current*, and is registered by a sensitive ammeter.

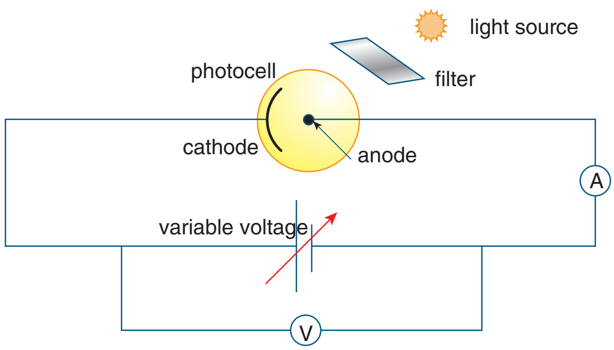


Figure 5.40

In 1916, Robert Millikan used apparatus similar to this to determine the threshold frequency, f_0, for a variety of metal cathodes. He found that for different metals a different minimum frequency was required in order to produce photoelectrons. He found that elements with higher ionisation energies had correspondingly higher values for the threshold frequency.

Physics file

At the end of the 19th century it was believed that light needed a medium in which to travel, called the 'aether'. The Americans Michelson and Morley had tried without success to find the speed of the Earth as it moved through this aether. Their experiment was designed to measure the speed of light from a star at two times in the year, 6 months apart. They expected to find a difference in the values they obtained since at one time, the Earth would have been travelling towards the star, and 6 months later the Earth would be on the other side of its orbit, and therefore travelling away from the star. However, Michelson and Morley obtained the same value for the speed of light: 3.00×10^8 m s^{-1} regardless of when they took their measurements. This was considered a strange result. Intuition would suggest that the speed of light from the star would be higher when the Earth was moving towards it, and lower as the Earth moved away. To explain the result, it took a tremendous revolution in mechanics in which Newton's mechanics had to be replaced by Einstein's theory of special relativity for fast moving bodies. Newton's work is now seen as part of a bigger, more encompassing theory.

Practical activity

27 Photoelectric effect

Interactive tutorial

Photoelectric effect: Investigate colour of light and the PE effect

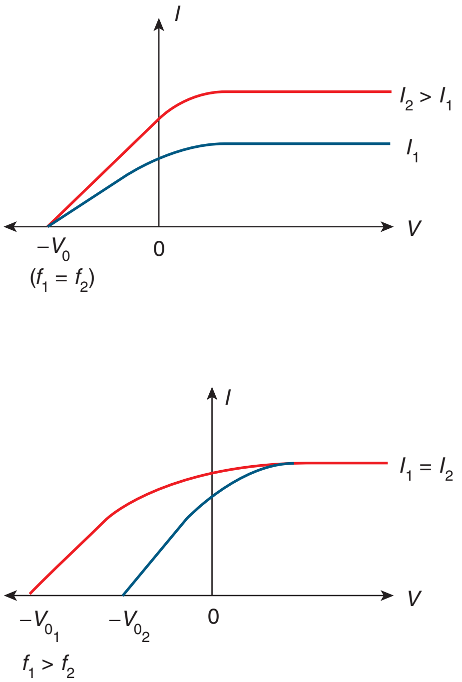


Figure 5.41

Photoelectric current plotted as a function of the applied voltage between the cathode and the anode in the photocell. A standard monochromatic light source where $f > f_0$ shows that with a forward potential, every available photoelectron is included in the current. With a reverse potential, the number of photoelectrons decreases until none are collected at the stopping voltage, $-V_0$. (a) For brighter light of the same frequency, there is a higher photoelectric current, but the same stopping voltage. (b) For light with a higher frequency, there is an increase in the stopping voltage.

Interactive tutorial

Photoelectric effect: Investigate light intensity

The circuit includes a variable voltage supply which can be used to make the cathode negative (and the **anode** positive). When this is done, the photoelectrons will be helped by the resulting electric field to the anode, and a maximum possible current will be measured. Alternatively, the voltage may be adjusted to make the cathode positive and the anode negative—a **reverse potential**. This arrangement is used to investigate the kinetic energy carried by the emitted photoelectrons.

The frequency of the light source can also be controlled with light filters. Filters which allow only a single frequency to pass make the light monochromatic, so the energy of the photoelectrons can be measured when light of different single frequencies is shining on the cathode. Finally, the cathode and anode are enclosed in an evacuated glass tube so that the photoelectrons do not suffer collisions with any gas molecules as they travel to the anode.

Using the arrangement in Figure 5.40, the German physicist Philipp Lenard found the following (for which he received a Nobel Prize in 1905).

- By varying the frequency of the incident light for a particular cathode metal, there is a certain frequency below which no photoelectrons are observed. This frequency is called the **threshold frequency**, f_0. For frequencies of light greater than the threshold frequency (i.e. $f > f_0$), photoelectrons will be collected at the anode and registered as a photoelectric current. For frequencies below the threshold frequency (i.e. $f < f_0$), no photoelectrons will be detected.
- For light whose frequency is greater than the threshold frequency, $f > f_0$, the rate at which the photoelectrons are produced varies in proportion with the **intensity** of the incident light (see Figure 5.41a). For frequencies below the threshold frequency, no photoelectrons are ejected no matter how intense the beam is made.
- As long as the incident light has a frequency above the threshold frequency of the cathode material, the ejected photoelectrons are found to be emitted without any appreciable time delay. This fact holds true regardless of the intensity of the light. In fact, modern experiments show any time delay to be as little as 10^{-9} s.

Stopping voltage and photoelectron energy

Lenard also used the apparatus to investigate the energy of the emitted photoelectrons. For this, he fixed the frequency of the incident light (above f_0, of course) and applied an increasing *negative* potential difference. As the reverse potential was increased from zero, the photoelectric current was seen to drop. This indicated that fewer and fewer photoelectrons had the energy to overcome the negative electric potential and reach the anode. At a certain fixed value, called the *stopping voltage*, V_0, no photoelectric current is registered.

Recall from our earlier studies of electricity that the work done on a charge (by an applied potential difference) is given by $W_d = q\Delta V$. In this case, the potential difference used is designated the stopping voltage, V_0, and the charge value is equal to the charge on an electron, e, -1.6×10^{-19} C. Hence the work done on the electron is given by $W_d = eV_0$. Since the stopping voltage is large enough to stop even the fastest moving electrons from reaching the anode, this expression gives the value of the maximum possible kinetic energy of the released electrons.

Lenard deduced that:

- The photoelectrons have a range of speeds up to a maximum speed.
- As the frequency of the incident light is increased, the maximum kinetic energy of the photoelectrons increases, as seen by an increase in the stopping voltage. Importantly, Lenard also showed that for a given metal, the stopping voltage depends only on the frequency of the incoming light, and is totally independent of the intensity of the light.

■

The photoelectron with the maximum speed has a ***KINETIC ENERGY*** (J) given by:

$$E_{k(max)} = \frac{1}{2}mv_{max}^{2} = eV_0$$

where m is the mass of the ***electron*** (9.11×10^{-31} kg), v_{max} is the maximum speed of the electron (m s^{-1}), V_0 is the negative stopping voltage (V) and e is the charge on the electron (-1.60×10^{-19} C).

Alternative unit for energy

The SI unit for work done is the joule (J). A joule is the quantity of energy that a coulomb of charge would gain after being moved through a potential of 1 V. However, when dealing with the very small energy values, such as those involved in the study of the photoelectric effect, another (non-SI) unit is often used. This unit is the **electronvolt** (eV), the amount of energy an electron gains on moving through a potential difference of 1 V. It is a tiny fraction of a joule. Its name is a little misleading—keep in mind that it is a unit of energy and not potential.

■

One ***ELECTRONVOLT*** (eV) represents the *energy* that a single electron would gain after being moved through a potential of 1 V.

Therefore, if an electron was accelerated through a potential difference of 100.0 V, it can automatically be stated that it has gained 100.0 eV of kinetic energy. The conversion factor between joules and electronvolts is 1.6×10^{-19}. This is the value of the charge on an electron, since 1 eV = qV = $1.6 \times 10^{-19} \times 1 = 1.6 \times 10^{-19}$ J. Hence:

■

$$1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$$

Electron energy values are also commonly given in keV and MeV.

With reference to the photoelectric effect, the use of the electronvolt as a convenient unit of energy results in a useful relationship. As stated, when working in joules the $E_{k(max)} = eV_0$. When working in electronvolts, $E_{k(max)}$ is given directly by the value of the stopping voltage in volts. For example, should a stopping voltage of 2.50 V be required, it can automatically be stated that the maximum kinetic energy of any electron is 2.50 eV.

Physics file

Since we are often interested in the maximum speed at which electrons travel, it may be worth rearranging the relationship

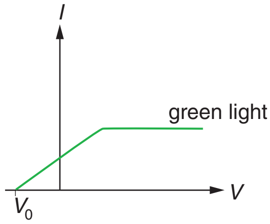
$$\frac{1}{2}mv_{max}^{2} = eV_0$$

so that the maximum speed is the subject, i.e.

$$v_{max} = \sqrt{\frac{2eV_0}{m}}$$

✓ Worked Example 5.4B

A sample of potassium is used as the cathode of a photocell with which the photoelectric effect is studied. When green light of a particular intensity is shone onto the cathode, the following *I*–*V* graph is obtained. Also, the threshold frequency for this sample is found to lie in the yellow region of the visible spectrum.



a What current reading would be expected if red light was shone onto the cathode?

b Draw the *I*–*V* graph that would result if the intensity of the incident green light was doubled.

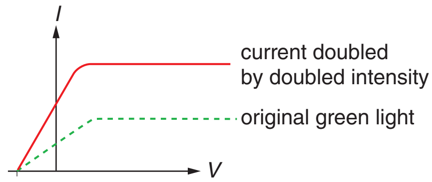
c Draw the *I*–*V* graph that would result if violet light of a very low intensity was incident upon the cathode.

d When UV light is incident upon the cathode, the stopping voltage is found to be –2.25 V. Determine the maximum kinetic energy of the photoelectrons in *joules* and *electronvolts*.

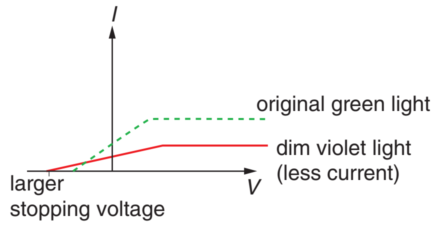
Solution

a Since red light corresponds to light of a lesser frequency than the threshold frequency (yellow light), no current will be observed.

b



c



d Working in joules:

$e = -1.60 \times 10^{-19}$ C

$V_o = -2.25$ V

$E_{k(max)} = eV_o$

$= (-1.60 \times 10^{-19})(-2.25)$

$= 3.60 \times 10^{-19}$ J

Working in electronvolts, by definition with a stopping voltage of –2.25 V acting on an electron:

$E_{k(max)} = 2.25$ eV

Note that the negative value of the stopping voltage is disregarded as we are finding the kinetic energy that the electron has, as opposed to what work must be done to remove the kinetic energy and stop the electron.

Wave model inadequate!

By Lenard's time, the wave model of light, as fully elaborated in Maxwell's equations for electromagnetic radiation, was taken as an article of faith. However, it was quickly seen that, no matter how the wave theory might be modified, it could not be used to explain the photoelectric effect.

In a wave model for light, a beam of light can be considered to be a series of wavefronts which will arrive at the metal surface with each wavefront simultaneously acting over the whole surface. The electric field component within each light wave would then start to cause any free electrons to vibrate. After a time, the electrons might have enough energy to escape. The following are several predictions that seem plausible if light behaved only as an electromagnetic wave.

- The wave model predicts that light of any frequency should produce photoelectrons. All light contains an oscillating electric field and, regardless of frequency, all light should be able to pry electrons free if the light source is of sufficient intensity. This is not seen. There is a frequency for light, called the threshold, f_0, below which no photoelectrons will be emitted.
- The wave model for light assumes that energy *E* delivered to the electrons in the metal will be given by the relationship $E = P\Delta t$, where *P* is the power of the beam and Δt is the time over which the surface is illuminated. Since the light beam will interact simultaneously across the whole metal surface, the energy of the beam will be divided

equally among all the electrons. This would suggest that if the beam has a low intensity, then photoelectrons will be emitted, but after a longer time, since it should take some time before sufficient energy is delivered to the electron. However, no time delay is evident, even with a very low-intensity source. There may not be a large photoelectric current, but the emission is seen to be instantaneous.

- Finally, using a wave model one would expect that an intense beam will cause the photoelectrons to be emitted with a greater kinetic energy. An intense beam would provide more power, so the energy delivered to an electron would be greater. What is seen is that increased beam intensity means more photoelectrons of the same energy. The photoelectric current is seen to be directly proportional to the intensity of the beam, but electron energy is not affected by the light intensity.

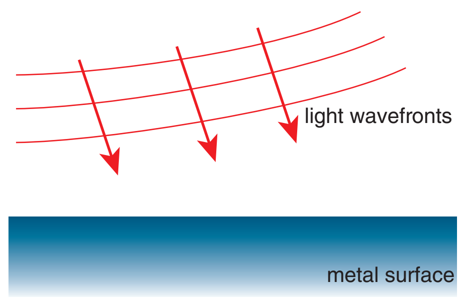


Figure 5.42
The wave model predicts that a wavefront will interact with all electrons near the surface of a material at once. The total energy available will be divided among the electrons. If the light beam has a low intensity, the wave model predicts that there will therefore be a time delay before any electrons are ejected from the metal.

Photons—a new model for light?

In 1900, the German physicist Max Planck was attempting to explain the spectrum for the light emitted by a hot object (called a black-body radiator). To do this, he developed the idea that light energy is not *continuous* (as in the wave model), but is delivered in tiny, discrete bundles called **photons** or *quanta*. He suggested that the energy carried by each photon is proportional to the frequency of the light.

The energy, E, carried by each ***PHOTON*** of light depends on the frequency, f, of the light so that:

$$E_{ph} = hf \quad \text{and} \quad f = \frac{c}{\lambda}$$

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

where h is Planck's constant (6.63×10^{-34} J s), f is the frequency of the photon (Hz), c is the speed of light (3.00×10^{8} m s^{-1}), λ is the wavelength (m) and E_{ph} is the photon energy (J).

This was such a revolutionary idea that Planck was widely ridiculed in scientific circles and his contemporaries spent the next few years trying to disprove his suggestion. At its heart, the idea of photons suggested a *particle model* for light. Since the predictions about the interaction between light and matter were distinctly different, depending on whether a wave or a particle model was used, this new thinking was threatening to many older physicists.

Planck's photon model suggests that the energy carried by a beam of light consists of a number of discrete packets of light energy (the photons). This means that the total energy carried by the beam will be $N \times hf$, where the energy carried by each photon will be hf, and the number of photons in the beam is N. Furthermore, Planck suggested that a single photon can only interact with one electron in the material at a time. If a photon is absorbed by a particle such as an electron, it will be completely absorbed, transferring all of its energy at once.

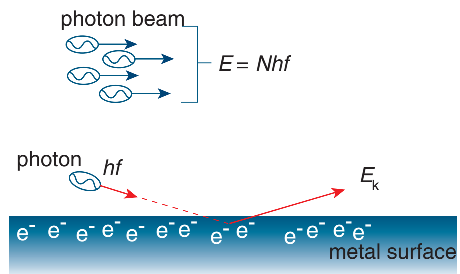


Figure 5.43
The photon model for light suggests that the beam of light carries its energy as a stream of photons, and that each photon interacts with one electron in the metal. A low-intensity beam simply has fewer photons, but each photon still carries the same energy.

Physics file

It is often convenient to give the energy of a photon in the non-SI unit, the *electronvolt*. As 1 eV = 1.60×10^{-19} J, the value of Planck's constant, h, can be expressed as:

$$h = \frac{6.63 \times 10^{-34}}{1.6 \times 10^{-19}}$$

$$= 4.14 \times 10^{-15} \text{ eV s}$$

$E = hf$ still applies but, using the alternative value of Planck's constant, the energy will now be in electronvolts.

✓ Worked Example 5.4C

A 100.0 W light globe produces yellow-green light of wavelength 500.0 nm. Determine the number of photons released from the globe every minute.

Solution

The globe will release a total energy of:

$P = 100.0$ W

$\Delta t = 60.0$ s

$E = P\Delta t = (100.0)(60.0)$
$= 6.00 \times 10^3$ J

The frequency of the photon is:

$c = 3.00 \times 10^8$ m s^{-1}

$\lambda = 500.0 \times 10^{-9}$ m

$f = \frac{c}{\lambda} = \frac{3.00 \times 10^8}{500.0 \times 10^{-9}}$
$= 6.00 \times 10^{14}$ Hz

This energy is carried by N photons, and the energy of each photon will be given by:

$E = 6.00 \times 10^3$ J

$f = 6.00 \times 10^{14}$ Hz

$h = 6.63 \times 10^{-34}$ J s

$E = Nhf$

$N = \frac{E}{hf} = \frac{6.00 \times 10^3}{(6.63 \times 10^{-34})(6.00 \times 10^{14})}$
$= 1.51 \times 10^{22}$ photons

1.51×10^{22} photons emitted each minute—a very large number indeed!

The dual nature of light

Einstein's explanation of the photoelectric effect

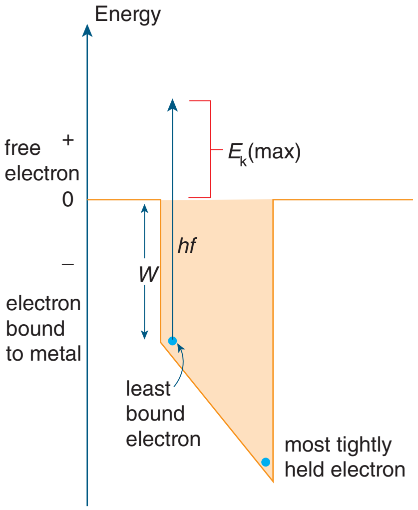


Figure 5.44

Electrons are bound by different amounts of energy. Photons with sufficient energy will free the least bound electron with the greatest speed.

In order to explain the photoelectric effect, Einstein used the photon concept that Planck had developed. He considered that, within the metal, each electron was bound to the metal by a different amount of energy. (At that time he knew nothing of electron shells.) Some electrons required substantial amounts of energy to become free, while others required less energy. Einstein was able to represent this situation using a 'potential well'. If the y-axis represents the total energy of the electrons, the electron will be bound to the metal where the energy is negative, will be freed where the energy is positive, and this energy is kinetic energy. An electron with zero energy would be free, but have no speed.

When the frequency of the incident light is less than the threshold frequency ($f < f_0$), the extra energy gained by even the least bound of the electrons is not enough for it to be freed. This is why no photoelectrons will be emitted for incident light of frequency lower than the threshold frequency, no matter how intense it is. For frequencies greater than the threshold frequency ($f > f_0$), the absorption of a photon can free some electrons. The electron that escapes with the greatest kinetic energy (and hence speed) will be the least bound electron.

The minimum energy required to release an electron from the metal is called the **work function** (W) for the metal. This equates to the situation where a photon at the threshold frequency is absorbed by the least bound electron. For this case, the ejected photoelectron will escape with no kinetic energy.

The minimum energy required to release an electron, the ***WORK FUNCTION*** (W), is given by:

$$W = hf_0$$

where h is Planck's constant (6.63×10^{-34} J s or 4.14×10^{-15} eV s) and f_0 is the threshold frequency (Hz).

By considering only the least bound electron, Einstein was able to provide a mathematical treatment. If light whose frequency is greater than the threshold frequency ($f > f_0$) shines on the metal in question, then a least bound electron absorbing the photon will be released with the largest kinetic energy of any photoelectron. Using the **law of conservation of energy**, all the energy of the photon, hf, will be passed to the electron. Some of this energy, the work function, W, will enable the electron to escape from the metal, and the remainder will be converted to kinetic energy for the electron, $E_{k(max)}$. So:

$$hf = W + E_{k(max)}$$

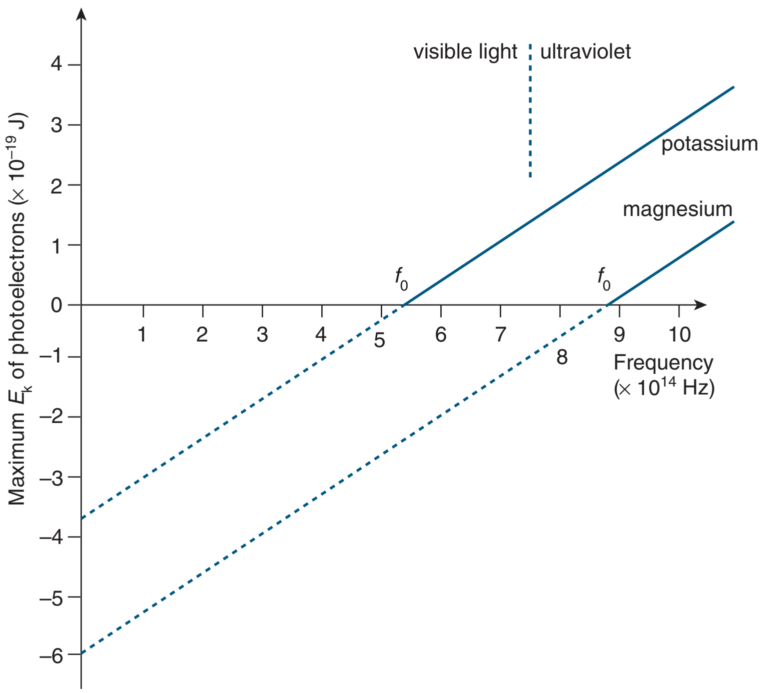
Rearranging, the maximum kinetic energy for the least bound electron becomes:

$$E_{k(max)} = \frac{1}{2}mv^2_{max} = hf - W$$

where $E_{k\ (max)}$ is the maximum kinetic energy of the released electrons (J), hf is the incident photon energy (J) and W is the work function (J). Note that eV or J may be used as the units of energy, but only if each variable has the same unit.

Experimentally, the maximum kinetic energy is found from the stopping voltage, V_0, since this is the smallest negative potential whose electric field will prevent even the fastest photoelectron (i.e. least bound electron) from reaching the anode. As we have discussed earlier in this chapter:

$$eV_0 = \tfrac{1}{2}mv_{max}^{\ 2}$$



Physics file

By comparison of the form of the equation for a straight line:

$y = mx + c$

and Einstein's finding:

$E_{k(max)} = h\ f - W$

↑ ↑↑ ↑

$y = m\,x + c$

it can be seen that the slope of the graph of $E_{k(max)}$ vs. f is Planck's constant and by extrapolating back to the y-axis will give the magnitude of the work function, W.

Interactive tutorial

Photoelectric effect: Investigate frequency and maximum E_k

Figure 5.45

Magnesium has a high threshold frequency, which is in the ultraviolet region. The threshold frequency for potassium is in the visible region. The gradient for the graph for each metal will be Planck's constant, h.

✓ Worked Example 5.4D

Yellow-green light of wavelength 500.0 nm shines on a metal whose stopping voltage is found to be –0.800 V. The mass of an electron is 9.11×10^{-31} kg. Find the:

a speed of the fastest moving photoelectron produced

b work function of the metal in both joules and electronvolts.

Solution

a $e = -1.60 \times 10^{-19}$ C

$V_o = -0.800$ V

$m = 9.11 \times 10^{-31}$ kg

$$E_{k(max)} = eV_o + \frac{1}{2}mv^2_{max}$$

$$v_{max} = \sqrt{\frac{2eV_o}{m}} = \sqrt{\frac{2(-1.60 \times 10^{-19})(-0.800)}{9.11 \times 10^{-31}}}$$

$$= 5.30 \times 10^5 \text{ m s}^{-1}$$

b $h = 6.67 \times 10^{-34}$ J s

$c = 3.00 \times 10^8$ m s^{-1}

$r = 500.0 \times 10^{-9}$ m

$e = -1.60 \times 10^{-19}$ C

$V_o = -0.800$ V

$$E_{k(max)} = hf - W$$

$$W = hf - E_{k(max)}$$

$$= \frac{hc}{r} - eV_o$$

$$= \frac{(6.67 \times 10^{-34})(3.00 \times 10^8)}{500.0 \times 10^{-9}} - (-1.60 \times 10^{-19})(-0.800)$$

$$= (4.00 \times 10^{-19}) - (1.28 \times 10^{-19})$$

$$= 2.72 \times 10^{-19} \text{ J}$$

The work function is 2.72×10^{-19} J:

$$W = \frac{2.72 \times 10^{-19}}{1.60 \times 10^{-19}}$$

$$= 1.70 \text{ eV}$$

Physics file

When applying the wave equation $v = f\lambda$ to any form of EMR travelling in air, do not forget that $v = c$ Therefore, instead of:

$$f = \frac{v}{\lambda}$$

we say:

$$f = \frac{c}{\lambda}$$

As established by Einstein's equation, the graph is a straight line whose gradient is Planck's constant. The *y*-intercept gives a value of the work function for the metal in the cathode. Different types of metal cathode will have different threshold frequencies, but all will produce graphs with the same gradient.

Momentum of the photon

In 1921 Einstein received the Nobel Prize for his work on the photoelectric effect and the theory of relativity. At the time, many physicists reasoned that, if light could carry energy just as a particle does, it might also carry momentum.

Direct evidence for this came in 1923 when Arthur Compton aimed a monochromatic beam of X-rays at a small block of graphite. Compton found that X-rays emerged at all angles from the block. He found that, at each angle, there were scattered X-rays of two X-ray wavelengths. One had the same value as the incident X-rays, and the other had a longer wavelength. The X-rays emerging from the block with an identical wavelength to the incident X-rays were considered to have bounced off the carbon atoms (like a ping-pong ball from a bowling ball) with their energy unaltered.

The wave model accounted for the X-rays of identical wavelength, but it could not account for those with the longer wavelength. A photon perspective was needed. Using this framework, the incident X-ray photons had lost some of their energy to an electron ejected from the graphite. Compton found that the wavelength of these scattered X-rays varied with the scattering angle, which prompted him to investigate the photon from the point of view of its momentum (a vector quantity).

Compton considered these X-ray interactions to be elastic, and similar to collisions between billiard balls, where energy and linear momentum need to be conserved. Compton equated the energy of a photon to Einstein's mass–energy equivalent of the photon hence:

The ***MOMENTUM*** of a photon, $\boldsymbol{p}$, is given by:

$$E = hf \quad \text{and} \quad E = mc^2$$

$$mc^2 = hf$$

$$\frac{mc^2}{c} = \frac{hf}{c}$$

$$mc = h\frac{f}{c}$$

$$= \frac{h}{\lambda}$$

Compton interpreted the product 'mc' to be the **momentum of the photon**, so:

$$mc = \boldsymbol{p} = \frac{h}{\lambda}$$

where $\boldsymbol{p}$ is the momentum of the photon (kg m s^{-1}), h is Planck's constant (6.67×10^{-34} J s) and λ is the wavelength of the photon (m). Note that h must be applied in its SI unit, $h = 6.63 \times 10^{-34}$ J s.

By adding the property of momentum to the photon, a particle nature for light would seem to be indisputable. Interactions between light and matter could now be seen as two particles interacting with each other. In fact, Compton referred to the X-rays in his experiment as having been scattered either elastically or inelastically—terms which previously had only related to matter.

The dual nature of light

By the middle of the 1920s, this ambiguous picture for the nature of light had been confirmed. Light has a dual nature: it behaves like both a wave and a particle. Each view is well supported with experimental evidence, and each experiment can only be explained by one model. If we adopt a wave model, we cannot explain the photoelectric effect and Compton scattering. If we treat light as a photon, we cannot explain phenomena such as interference and diffraction. Significantly, no experiment has yet been devised in which light displays both its natures simultaneously. Sir William Bragg summed up the problem this way: 'On Mondays, Wednesdays and Fridays, light behaves like waves, and on Tuesdays, Thursdays and Saturdays like particles, and like nothing at all on Sundays.'

In an interference experiment, the alternating bright and dark fringes were once described as regions where the light waves have reinforced or cancelled. When light is considered as photons, we can only interpret the interference effects as dictating where the photons are to go.

Individual photons display wave property

In 1909 G. I. Taylor attempted an interesting interference experiment in which light was forced to behave only as a photon. A very weak source of light was directed at the two slits of a Young's interferometer.

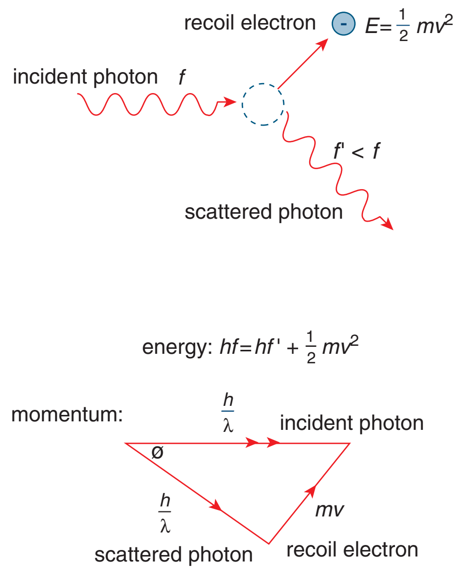


Figure 5.46
Compton scattering involves X-ray photons of a given wavelength giving both energy and momentum to a recoil electron. The collision can be considered in exactly the same way as one might analyse two massive objects interacting. In some situations, the scattering will be elastic, where no photon energy is transferred to the electron. Other collisions will be inelastic: the electron will be given some energy and momentum from the X-ray photon.

Physics file

The discussion of the momentum of a photon presented here is overly simplified. Realise that, since a photon travels at the speed of light, a relativistic approach should be taken when deriving expressions regarding its momentum and energy. However, since a photon has zero rest mass its energy expression reduces to $E = mc^2$ anyway.

The source was so weak that only one photon could be present between the slits and the screen at any one time. There was absolutely no chance for interference to occur! To view the pattern on the screen, Taylor had to use a photographic plate, and he left the experiment to run for 3 months so that the film would be adequately exposed.

The result of Taylor's experiment was a set of interference fringes, just as if a more intense light had been used over a shorter period. How could this be? Only single photons were used: what was their path to the screen? If they could not interfere, how did they know to leave some spaces?

The interference pattern is providing only certain pathways for the photons, and we interpret the intensity of interference pattern as a probability distribution for the photons. Where there is a bright fringe, there is a high probability that the photons will land there.

The difficulty we have in trying to create a single model for light is that light is not something we can handle or look at directly. The best we can do is to try to picture what is occurring by using a mental model based on experimental evidence. These models are further tested by experiment to see how well they perform.

Light really cannot be thought of as a simple wave or particle. It is just light! However, as humans, we want to understand it by relating it to things about us of which we have a good understanding. We can use part of the ideas we associate with particles, add in some wave ideas and probability theories and begin to accept light for its subtlety and complexity.

5.4 SUMMARY Electromagnetic radiation

- The photoelectric effect is the emission of photoelectrons from a clean metal surface due to incident light whose frequency is greater than a threshold frequency, f_0.
- If $f < f_0$, no electrons are released. If $f \geq f_0$, the *rate* of electron release (current) is proportional to the intensity of the light and occurs without any time delay.
- The reverse voltage can be increased until it is large enough to stop even the most energetic electrons from reaching the anode. Thus the stopping voltage, V_0, indicates the maximum kinetic energy of the photoelectrons, E_k(max).
- The wave approach to light could not explain various features of the photoelectric effect: the existence of a threshold frequency, the absence of a time delay when using very weak light sources, and increased intensity of light resulting in a greater rate of electron release rather than increased electron energy.
- Max Planck developed a photon model for light. The energy carried by each photon is given by: $E_{photon} = hf$, where h = Planck's constant $= 6.63 \times 10^{-34}$ J s or 4.14×10^{-15} eV s.
- Einstein used Planck's concept of a photon to explain the photoelectric effect, stating that each electron release was due to an interaction with only one photon.
- The photon approach explained the existence of a threshold frequency for each metal, the absence of a time delay for weak light sources and why brighter light resulted in a higher photocurrent.
- The work function, W, for the metal is given by $W = hf_0$, and is different for each metal. If the frequency of the incident light is greater than the threshold frequency, then a photoelectron will be ejected with some kinetic energy up to a maximum value.
- $E_{k\,(max)}$ (J) $= eV_0$, where V_0 is the stopping voltage.
- The momentum of a photon, $\boldsymbol{p}$, is given by: $\boldsymbol{p} = \frac{h}{\lambda} = \frac{hf}{c}$ where $h = 6.63 \times 10^{-34}$ J s.
- As a consequence of the explanation of the photoelectric effect and the allocation of the property of momentum, we now understand that light has a dual nature—wave-like and particle-like. The photon model suggests that light energy is quantised rather than being continuous.

5.4 Questions

1 a List three different types of electromagnetic radiation and describe a use for each.
b List two properties common to all forms of electromagnetic radiation.

For the following questions use:

Planck's constant = 6.63×10^{-34} J s or 4.14×10^{-15} eV s

speed of light = 3.00×10^{8} m s^{-1}

charge on electron = -1.60×10^{-19} C

mass of electron = 9.11×10^{-31} kg

2 The frequency of a green light is 5.60×10^{14} Hz.
a What is the wavelength of this light?
b Calculate the energy of a photon of this green light in joules and electronvolts.

3 Which of the following statement(s) are true with respect to the value of the stopping voltage obtained when using a photocell?
A The stopping voltage indicates how much work must be done to stop the most energetic photoelectrons.
B The stopping voltage is proportional to the square of the speed of the fastest electrons.
C The stopping voltage is associated with a situation in which there is also no photocurrent.
D If only the brightness of the incident light is increased, the stopping voltage will not alter.
E For a given metal, the value of the stopping voltage is affected only by the frequency of the incident light.

4 The stopping voltage that is obtained using a particular photocell is 1.95 V.
a Determine the maximum kinetic energy of the photoelectrons in:
i electronvolts
ii joules.
b Determine the speed of the fastest photoelectron.

5 Which of the following is correct? Increasing the brightness of a source of yellow light will:
A increase the energy of the photons emitted from the source
B reduce the wavelength of the photons emitted from the source
C increase the number of photons emitted per second from the source
D increase the wavelength of the photons emitted from the source.

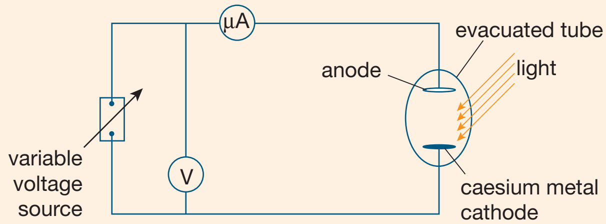
6 Confirm whether each of the following statements relating to the photoelectric effect and the photon model is true or false.
a A bright light source emits the same number of photons as a dim light source, but each photon has more energy.
b For a given sample of metal, light of sufficient energy can produce the photoelectric effect regardless of the intensity.
c For a given sample of metal, light of sufficient intensity can produce the photoelectric effect regardless of the energy.
d The number of photoelectrons produced is determined by the intensity of the incident light.

7 Two separate photocells are set up using two light sources of different intensities and frequencies of light. Identical metal cathodes are used and the photoelectric effect occurs in each cell. Which one or more of the following findings will be made?
A The stopping voltage and photocurrent will be identical in each cell.
B The more intense light will require a larger stopping voltage.
C The more intense light will result in a larger work function value.
D The more intense light will result in a larger maximum kinetic energy of the photoelectrons.
E The more intense light will result in a larger photocurrent.

8 Two separate photocells are set up using two light sources of the same intensity and frequency of light. Different metal cathodes are used in each cell, each having a different known work function. The photoelectric effect is observed in both cells. Which one or more of the following findings will be made?
A The stopping voltage will be identical in each cell.
B The photocurrent will be identical in each cell.
C The metal with the larger work function will require a larger stopping voltage.
D The metal with the larger work function will result in a larger maximum kinetic energy of the photoelectrons.
E The stopping voltage will be larger for the metal with the smaller work function.

The following information applies to questions 9 and 10.

A photoelectric cell is connected into a circuit. The longest wavelength of light that will eject electrons from a caesium surface is 652 nm. Orange light of wavelength 620 nm is incident on the caesium surface.



9 a What is the threshold frequency for caesium?
b Calculate the work function for caesium in electronvolts.
c What is the energy, in eV, carried by the incident photons of orange light?
d Find the kinetic energy in eV of the fastest photoelectrons emitted from the caesium surface.
e What is the momentum of the fastest photoelectrons emitted from the caesium surface?

10 The variable voltage source can be adjusted so that the anode can be made negative with respect to the cathode. The minimum retarding potential difference required to produce zero current in the circuit is V_o.

a Calculate the stopping voltage when orange light is incident on the cathode.

b Explain why no current flows in the circuit when this minimum retarding potential difference is applied between the cathode and anode.

c Yellow light of frequency 5.20×10^{14} Hz is incident on the cathode. Calculate the stopping voltage.

5.5 Electromagnetic radiation and matter

In section 5.4 the quantum expression for the energy of a photon was introduced. The work done by Planck and later Einstein not only led to the photon model for light, but also laid a foundation upon which the structure of the atom itself could begin to be understood. Once physicists had discovered that light of a given frequency carried a very specific quantum of energy, as described by the equation $E = hf$, the light that for many years had been observed being absorbed and emitted by atoms suddenly provided information about the energy state of the atom itself.

The old model of the atom

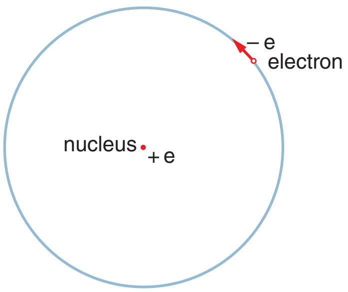


Figure 5.47
Rutherford's nuclear model of the hydrogen atom of 1911 did not explain the atom's known stability.

Practical activity
53 Spectra of different elements

In the early 20th century, the *nuclear model* of the atom was established through Ernest Rutherford's work in shooting alpha particles through gold foil. In the nuclear model, 99.9% of the mass of the atom is concentrated in the nucleus. The radius of the nucleus was around 10^{-15} m but the radius of the atom itself was some 40 000 times this value. The nucleus was known to carry a positive charge and by atomic dimensions was a tiny but massive dot in the centre of the atom. Sufficient electrons to balance the charge on the atom were thought to be floating somehow outside the nucleus. It was known that the simplest of all atoms was the atom of hydrogen, with one proton in the nucleus and one electron orbiting it.

Rutherford's model was not without its problems. An electron orbiting in this manner, according to the laws of the day, should gradually diminish in energy and spiral downwards towards the nucleus, emitting a range of energies as it went. This atom could not be stable and should continually give off energy. The hydrogen atom, of course, did not do this.

Observing emission spectra

During the late 1800s, scientists had devised methods of making atoms emit light. These methods involved heating substances, applying voltages to gases or indeed burning salts (Figure 5.48). The emitted light could be divided up into its different component frequencies by a spectrograph. Each type of atom was found to be capable of emitting a unique set of frequencies, its *emission spectrum* (Figure 5.49). Thus the emission spectrum of an atom became an individualised property of the atom.

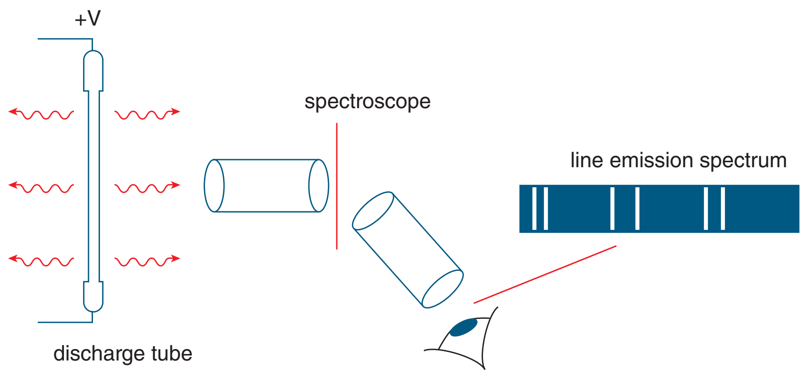
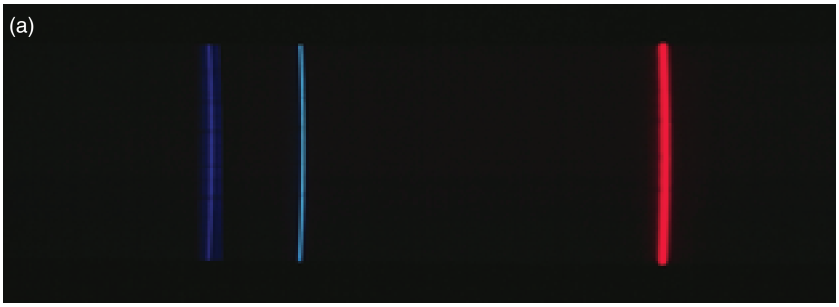


Figure 5.48
Placing gas into a tube and applying a high voltage to it is one technique for producing an emission spectrum. Sodium and mercury are two common examples of metal-vapour lamps.



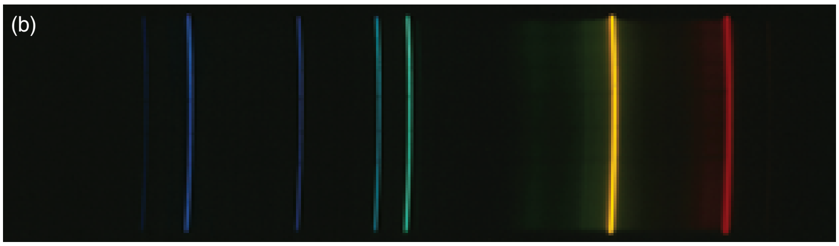


Figure 5.49
The emission spectra of (a) hydrogen and (b) helium, showing only the principal lines.

We saw earlier that the energy of a photon can be determined from a measurement of the photon's frequency or its wavelength according to the relationship:

PHOTON ENERGY is given by:

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

where E is the photon energy (J or eV), h is Planck's constant (6.63×10^{-34} J s or 4.14×10^{-15} eV s), c is the speed of light (3.0×10^{8} m s^{-1}) and λ is the wavelength (m).

Hydrogen atoms produce the simplest spectrum. Figure 5.49 shows a series of spectral lines, with each line corresponding to a particular frequency, and therefore energy, of light. Even this simplest of atoms was found to emit up to 40 discernible frequencies of light. Although physicists could not *explain* their existence, they could list and describe them in detail. The spectral lines seemed to have a mathematical pattern or order. They occurred in groupings, and in each group the higher the frequency the closer the spectral lines occurred, until infinitely closely spaced lines approached a limiting frequency value. Gradually it was found that three related mathematical expressions described the three sets of frequencies emitted by hydrogen (see Figure 5.50). As technology improved, more sets of spectral lines were discovered. It was many years before a meaningful interpretation of the emission spectrum was attained, but eventually, by utilising the emerging knowledge of the photon, the emission spectrum provided clues about the atom's structure.

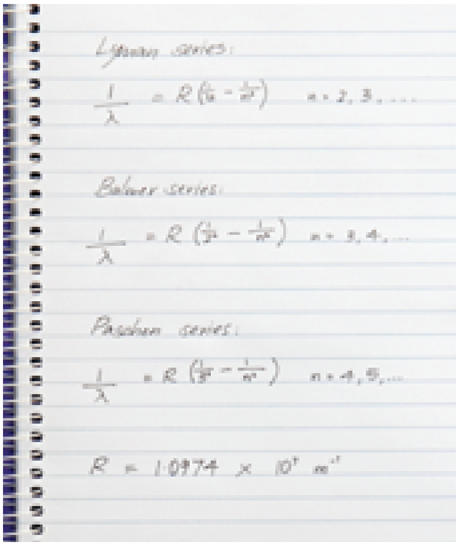


Figure 5.50
Years before emission spectra were understood, mathematical descriptions of the series of frequency values for hydrogen were devised.

Observing absorption spectra

Recall that white light is actually made up of an infinite number of different frequencies of light. If white light is passed through a spectroscope, a continuous rainbow of colours is seen, containing all the shades of the spectrum from red to violet. If infrared or ultraviolet frequencies are present, they too may be detected. When a beam of white light is passed through a cool sample of a gas, an *absorption spectrum* is created. The light that has passed through the gas is passed into a spectroscope as shown in Figure 5.51. Instead of giving a continuous spectrum, this shows that particular frequencies of light are missing. There are black spaces in locations throughout the spectrum, indicating that some frequencies have been *removed by the gas*. Like the emission spectrum, the absorption spectrum is once again *unique to each type of atom*. Furthermore, for a given sample, the missing lines in its absorption spectrum correspond exactly to bright lines in its emission spectrum; however, the emission spectrum also includes many more lines.

Practical activity

26 Wavelength of LEDs

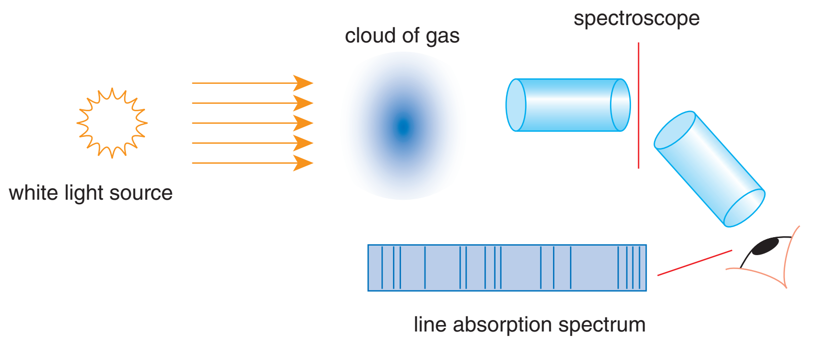


Figure 5.51
A line absorption spectrum is produced when white light passes through a cold sample of the material and is viewed through a spectroscope.

Physics file

Astronomers became interested in absorption spectra, and in 1862 Jonas Ångström (a Swede) was able to demonstrate the presence of hydrogen in the Sun from the solar spectrum. In 1868 it was discovered that there were lines in the solar spectrum that no known element on Earth could produce. Helium had been discovered in the Sun—30 years before it was found on Earth. (The name 'helium' is derived from the Greek *helios*, which means 'sun'.)

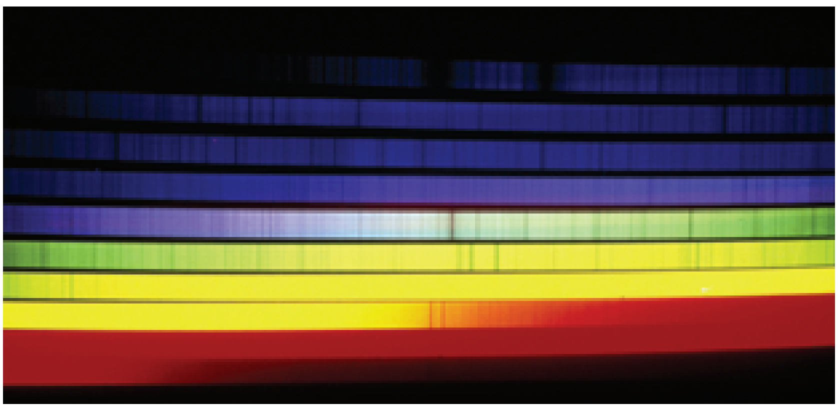


Figure 5.52
The solar spectrum as seen at the Earth. The dark absorption lines in the spectrum are caused by particular frequencies of light being absorbed by cool gases surrounding the Sun.

Bohr model for emission and absorption spectra

In 1912 Niels Bohr proposed the first reasonable interpretation of the hydrogen spectrum. The existence of absorption and emission spectra for hydrogen, together with Planck's quantum energy relation, $E = hf$, provided some starting points for Bohr. He realised that:

- absorption spectra showed that the hydrogen atom was only capable of absorbing a small number of different frequencies, and therefore energies, of very specific values; that is, the absorbed energy was *quantised*
- the emission spectra showed that hydrogen was also capable of emitting quanta of the *exact energy* value that it was able to absorb
- if the frequency, and therefore energy, of the incident light were below a certain value for the hydrogen atom, the light would simply *pass straight through* the gas without any absorption occurring
- hydrogen atoms have an *ionisation energy* of 13.6 eV; light of this energy or greater can remove an electron from the hydrogen atom, leaving it a positive ion
- all of the photons of light with energies above the ionisation energy value for hydrogen are continuously absorbed.

Starting with the nuclear model, Bohr chose to ignore its associated problem with energy emission yet sustained stability. He just presumed that for some unknown reason the atom was stable. (De Broglie later provided a neat explanation for such an assumption) Moving on from this assumption, Bohr devised a sophisticated model of electron energy levels for the atom, work for which he was later awarded a Nobel Prize in Physics.

Bohr's main ideas were as follows.

- The electron moves in a circular orbit around the nucleus of the hydrogen atom.
- The centripetal force keeping the electron moving in a circle is the electrostatic force of attraction (positive nucleus attracts negative electron).
- A number of allowable orbits of different radii exist for each atom and are labelled $n = 1, 2, 3, \ldots$ *The electron may occupy only these orbits.*
- An electron ordinarily occupies the lowest energy orbit available.
- The electron does not radiate energy while it is in a stable orbit.
- Electromagnetic radiation can be absorbed by an atom when its photon energy is *exactly equal to the difference in energies between an occupied orbit and a higher energy orbit*. This photon absorption results in an excited atom.
- Electromagnetic radiation is emitted by an excited atom when an electron falls from a higher energy level to a lower energy level. The photon energy will be exactly equal to the energy difference between the electron's initial and final levels.

■

If an electron in an atom moves between energy levels m and n, the energy of the photon that is either emitted or absorbed is given by:

$$E_{photon} = hf = E_m - E_n$$

where h is Planck's constant (6.63×10^{-34} J s) and f is the frequency of the photon (Hz).

Note that energy values are commonly quoted in eV, although joules or eV may be used as long as consistency is maintained.

Physics file

The discrete emission spectra studied in this chapter are a relatively uncommon form of emission spectrum. The EMR around us is commonly produced by the continual motion of atoms due to their thermal energy. EMR is produced whenever electric charges accelerate. Vibrating atoms result in the emission of a continuous range of frequencies simultaneously. The dominant frequency depends on the temperature of the object.

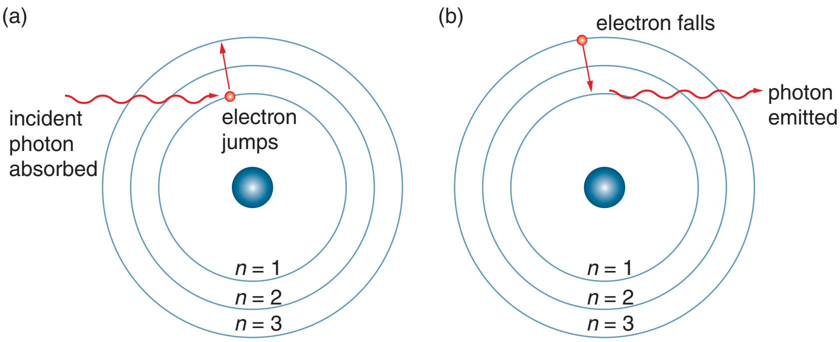


Figure 5.53

(a) If the incident photon carries a matching amount of energy, the electron can be knocked to one of the higher orbits as the photon is absorbed. The photon ceases to exist. (b) An atom will remain in an excited state for less than a millionth of a second. The electron will then fall to its ground state. The electron may fall in one step, or in a number of stages, emitting a photon(s) as it falls.

Physics file

In his explanation of atomic spectra, Bohr only allowed electrons to sit in orbits that satisfied the equation:

$$mvr = \frac{nh}{2}$$

where n is the energy level number. The product of the mass (m), velocity (v) and radius (r), called the angular momentum, must be equal to one of a series of fixed values involving Planck's constant. Bohr had no physical justification for why only these orbits could explain the spectra.

Physics file

There are two systems in use for labelling the energy levels of an atom. Sometimes the ground level ($n = 1$) is allocated 0 eV and therefore the higher levels have positive values. Alternatively, the ground state is allocated a negative value and the ionisation energy level ($n = \infty$) has a value of 0 eV.

And so the observed absorption and emission spectra have to a large part been accounted for by Bohr. The missing lines in *absorption spectra* correspond to the energies of light that a given atom is capable of absorbing due to the energy differences between its electron orbits. Only incident light that is carrying just the right amount of energy required to raise an electron to an allowed level can be absorbed.

The emission spectrum of an atom includes all of the absorption spectral lines plus more lines because these correspond to the energies emitted as the electron falls down through orbits either in *stages* or in one large fall—these transitions can occur across single levels or multiple levels. If the electron falls in stages, then photons corresponding to the *differences* between possible energy levels will be emitted. Since an electron ordinarily occupies the lowest energy orbit, there is little chance of an absorption occurring from a high orbit to an even higher orbit. Since more downward transitions are possible than upward transitions, there are more emission lines than absorption lines for hydrogen.

An electron ordinarily occupies the lowest energy orbit. Incident light carrying insufficient energy to raise an electron from this lowest energy level to the next level will be unable to be absorbed by the atom. This is why incident light below a certain energy value would simply *pass straight through* the hydrogen gas without any absorption occurring.

If light with greater energy than the ionisation energy of an atom is incident, then any light energy value may be absorbed. In this case, the excess photon energy simply translates to extra kinetic energy for the released electron (recall the photoelectric effect studied in section 5.4).

The spectra of metal-vapour lamps

Light sources that emit most of their radiant power in a few visible narrow spectral lines are very efficient light sources. A metal-vapour lamp is an example of a line source. It has two electrodes (the positive *anode* and the negative *cathode*) sealed in a quartz bulb. Inside the bulb is an atmosphere of argon gas at a relatively high pressure. When a high voltage is applied between the electrodes, an arc is struck which ionises some of the argon atoms. Positive ions are accelerated into the cathode, heating it up and freeing more electrons, and electrons are accelerated towards the anode. The accelerating electrons collide with other argon atoms, which are excited to higher energy states. When the electrons in the argon gas de-excite, it gives off a violet-bluish glow called a *glow discharge*. As the lamp heats up, a small amount of metal inside the bulb vaporises. The accelerated electrons can now excite these metal atoms into discrete higher energy levels. When these states de-excite back to lower levels, photons are emitted with wavelengths characteristic of the particular transitions.

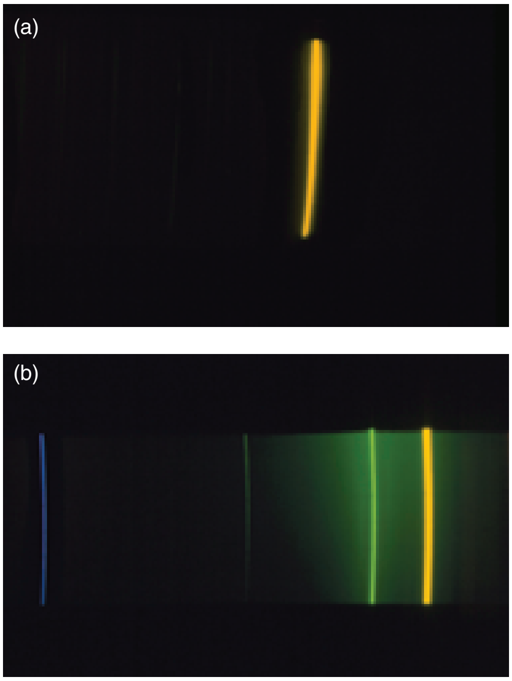


Figure 5.54
Emission spectra of (a) sodium and (b) mercury. The emission spectrum of a sodium- or mercury-vapour lamp consists of a large number of discrete frequencies, some of which lie in the visible region of the spectrum. Note the yellow line that gives sodium light its characteristic colour. Mercury's colours are blue, cyan, green and yellow.

Two common metal-vapour lamps use sodium and mercury. A sodium-vapour lamp generates line radiation at 589.0 and 589.6 nm (characteristic of the quantum atomic energy level transitions for the sodium atom). This lamp generates a yellow light. A mercury-vapour lamp generates line radiation at several wavelengths including 435.8, 546.1, 577.0 and 579.1 nm (characteristic of the quantum atomic energy level transitions for the mercury atom). These lines combine to generate a light dominated by the blue-green part of the visible spectrum. Sodium- and mercury-vapour lamps are commonly used in street lighting, because they produce a very efficient, bright light. They usually take about 10 minutes to warm up to their normal operating level.

Physics file

Some modern, and usually more expensive, cars have bluish headlights that are a variation of the mercury metal-vapour lamp. These lamps are called *metal-halide lamps*. In addition to the mercury spectral lines, they produce additional lines due to the presence of several exotic metal-halide salts. The result is a highly efficient bluish-white light. Xenon rather than argon is used to create the glow discharge, as this gives a brighter, more usable light during the warm-up stage.

Energy level diagrams

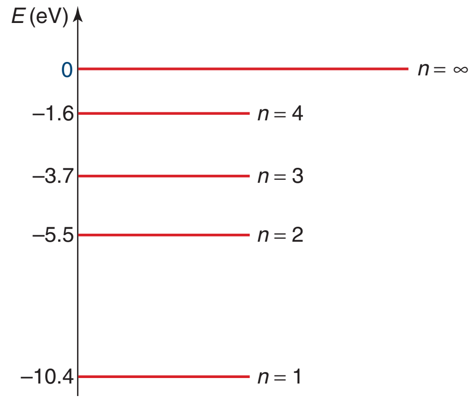
Energy level diagrams can be thought of as a section of the electron orbit diagrams shown in Figure 5.53, and a sample is shown in Worked example 5.5A. The allowed energy levels must be stated and these are written along the left-hand side of the diagram. Note that, since a free electron (at $n = \infty$) must possess zero potential energy, the energy levels within the atom are all negative. To raise an electron the appropriate amount of energy must be delivered by a photon. As an electron falls, its energy value decreases—that is, it becomes a larger negative number.

✓ Worked Example 5.5A

The energy levels for atomic mercury are depicted in the diagram.

a Consider the mercury atom with its valence electron in the ground state. Ultraviolet light with photon energies 4.9, 5.0 and 10.50 eV is incident on some mercury gas. What could happen?

b Determine the wavelength of the light emitted after an electron in an excited mercury atom makes the transition from $n = 3$ to $n = 1$.



Solution

a The 4.9 eV photon may be absorbed, promoting the electron from the ground state to the first excited state. The 5.0 eV photon cannot be absorbed since there is no energy level 5.0 eV above the ground state. The 10.5 eV photon may ionise the mercury atom. In this case, the ejected electron will leave the atom with 0.1 eV of kinetic energy.

b First determine the difference in the energy levels in joules:

$E_3 = -3.7$ eV

$E_1 = -10.4$ eV

$$\Delta E = (E_3 - E_1) \times (1.60 \times 10^{-19})$$
$$= [(-3.7) - (-10.4)] \times (1.60 \times 10^{-19})$$
$$= 1.07 \times 10^{-18} \text{ J}$$

Then calculate the wavelength of the emitted photons:

$h = 6.67 \times 10^{-34}$ J s

$c = 3.00 \times 10^{8}$ m s^{-1}

$\Delta E = 1.07 \times 10^{-18}$ J

$$E_{photon} = hf = \Delta E$$
$$h\frac{c}{\lambda} = \Delta E$$
$$\lambda = \frac{hc}{\Delta E} = \frac{(6.67 \times 10^{-34})(3.00 \times 10^{8})}{1.07 \times 10^{-18}}$$
$$= 1.87 \times 10^{-7} \text{ m}$$

λ = 187 nm (in the ultraviolet)

Band and continuous spectra

There are other types of spectra. Spectra can be classified as line, band or continuous. Band spectra are associated with molecules and polyatomic ions and continuous spectra are emitted from hot liquids and solids.

Band spectra

Band spectra differ from line spectra in that broad bands of frequencies are involved rather than sharp lines. The line spectra discussed in the previous section are characteristic of individual atoms in a gaseous phase while band spectra are characteristic of more complex substances such as those containing molecules or polyatomic ions. When band emission spectra are examined closely it can be seen that each band of colour comprises a number of closely spaced lines.

Band emission spectra are emitted from excited molecules in the gaseous state. Molecules have many more energy levels than atoms and the energy

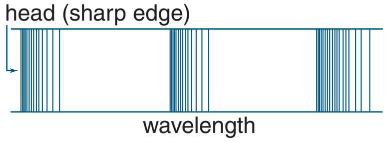


Figure 5.55
A typical band emission spectrum. Each band of colour is actually a large number of closely spaced lines. Typically, each band has a sharp edge where the lines are closely spaced and a fuzzy edge where the lines begin to spread out.

levels are in close groups. The extra energy levels result from overlap of the energy levels of the individual atoms in the molecule. Additional energy levels are also possible in molecules because they can rotate and the atoms can vibrate in relation to each other. This results in many more possible energy transitions, and consequently the emission spectra have a large number of lines. The lines are in close groups. In the early days of discovery poor resolution of equipment used for spectral analysis caused the lines to blend and form broad bands in the spectrum. Later it was discovered that the bands were made up of many lines but we still call them band spectra to distinguish them from the simpler line spectra of atoms.

Band absorption spectra result when white light is passed through coloured solids such as coloured glass or through a coloured solution, say, of potassium permanganate or copper sulfate. In these cases, the spectrum has some dark bands on the bright background of the continuous spectrum of white light being passed through the solid or the liquid. The brightness of some of the frequencies is decreased by absorption of some of the photons. This is why some solids and liquids have colours other than the primary colours of the rainbow. Their appearance is a result of small parts of the spectrum being missing rather than of viewing just a small part of the spectrum.

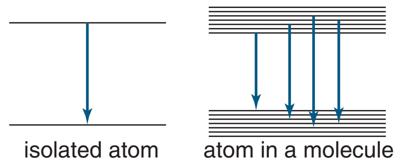


Figure 5.56

The individual energy levels of an isolated atom become bands of many closely spaced levels when the atom is in a molecule or in a body of solid or liquid.

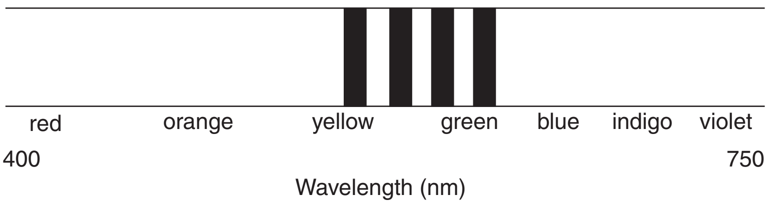


Figure 5.57

The absorption spectrum of a coloured solution. The black bands indicate which wavelengths have been absorbed by the solution.

Continuous emission spectra

Continuous spectra are emitted from hot solids such as the filament of an incandescent light globe or hot liquids such as molten iron from a smelter. In these substances, the atoms are so close together that their outer electrons overlap. The energy levels in the atoms are spread out into bands. The bands are very close together because of the very large number of atoms. Consequently so many different electron transitions are possible that the full range of wavelengths is produced.

This type of radiation is called black-body radiation and the specific characteristics of the continuous spectrum depend on the temperature of the emitting solid or liquid. At a temperature of around 1000 K a solid will glow 'red hot'. The filaments of light globes operate at around 3000 K and glow 'white hot'. At 3000 K other metals would be liquid but tungsten has a melting point of around 3700 K. This is why light globe filaments are made from tungsten. The light from incandescent light globes is rich in orange, red and infrared. At around 6000 K a black body emits a spectrum very much like the spectrum of the Sun. This is why photographs taken indoors under incandescent light often appear very red or orange. Most photographic colour film is designed for daylight conditions but it is possible to buy special colour film for 'tungsten lights'. Photographers often refer to the 'colour temperature' of a light source. This tells them about the colour effects they would expect from making photographs with that light source.

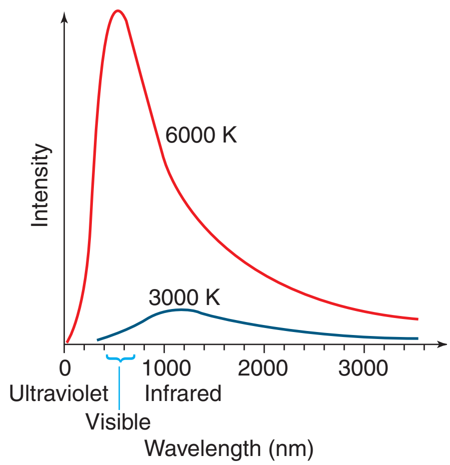


Figure 5.58

The continuous spectra emitted by a hot body at two different temperatures. At higher temperatures, the distribution includes a greater proportion of photons of higher energies and the total intensity is greater.

Fluorescence and phosphorescence

When an atom or molecule is excited to a higher energy state by photon absorption, it may return to its ground state in a series of two or more jumps. This typically happens within 10^{-8} s and emitted photons will have lower energies than the absorbed photon. This is called fluorescence and can result in line or band emission depending on the particular fluorescent material.

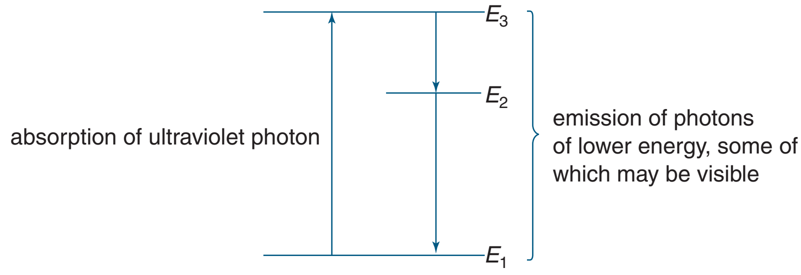


Figure 5.59

The process of fluorescence. An electron excited by the absorption of a high energy photon may then return to its ground state in more than one step, releasing photons of visible light.

Many fluorescent materials can absorb ultraviolet photons and then emit the energy as visible light photons. The colours emitted are characteristic of the fluorescent substance and consequently this phenomenon can be used for identifying substances. Sometimes a substance can be recognised by eye from its fluorescent spectrum and in other cases a spectrometer must be used. Another practical use of this phenomenon is the addition of fluorescent 'whiteners' to laundry detergent. These chemicals absorb ultraviolet light and re-emit the energy as blue light. This has the visual effect of brightening the fabric. Sometimes clothing will glow quite brightly under, so-called, black lights. 'Black lights' are lights designed to emit high levels of ultraviolet rather than visible light.

Some substances, when excited by photon absorption, form metastable states and don't re-emit the energy for some time. The electrons can take seconds or even hours to drop back to their ground state so light is emitted for some time. The process is identical to fluorescence but is called phosphorescence when there is a significant time delay involved. Phosphorescent substances are said to be luminous and are used in applications such as the paint for watch dials and the coatings on television screens.

Bohr's energy levels of hydrogen

Figure 5.60 shows the measured energy levels for hydrogen with its ground state of –13.6 eV. Each set of arrows indicates the energy level transitions that end at a common level. The energy level transitions to the ground state, called the Lyman series, produce a series of spectral lines in the ultraviolet region of the spectrum. The Balmer series includes energy transitions from various levels to the first energy level. The Paschen series ends at the second energy level. Bohr used classical physics ideas, involving the kinetic and potential energies resulting from the force between the orbiting electron and the positive hydrogen nucleus, to deduce that the value of the various allowed energy levels for the hydrogen atom could be represented by the equation:

$$E_n = \frac{-13.6}{n^2}$$

where E_n is the energy of the nth level for hydrogen (eV) and n is the energy level number 1, 2, 3, . . .
The negative sign indicates that energy must be added to the system to excite the atom.

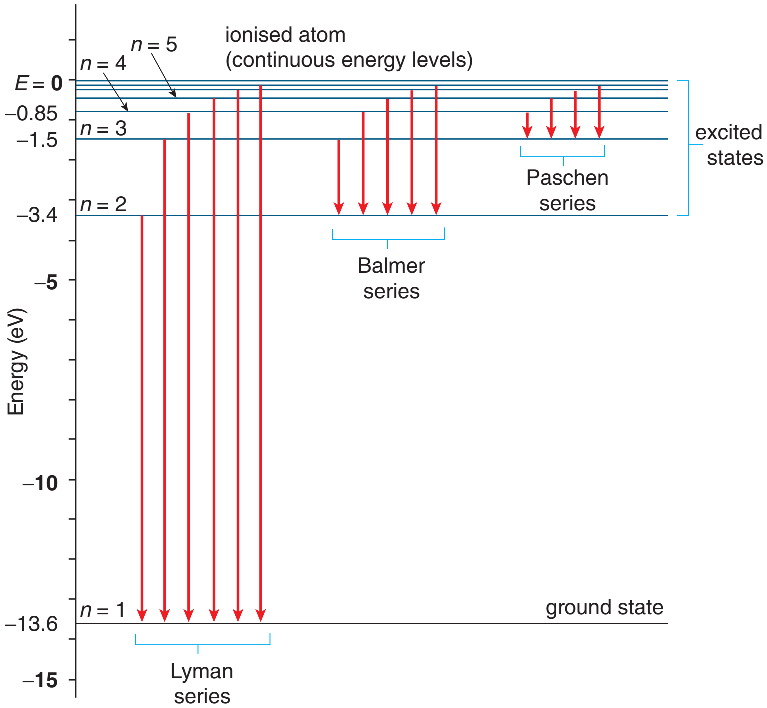


Figure 5.60
An energy level diagram for hydrogen. The ground state is bound by –13.6 eV and as one looks to higher energy levels, the energy levels are seen to crowd together.

Substituting in the known ionisation energy of 13.6 eV for hydrogen, Bohr obtained:

$$E_1 = \frac{-13.6}{1^2} = -13.6 \text{ eV (the ground state)}$$

$$E_2 = \frac{-13.6}{2^2} = -3.4 \text{ eV (the first excited state)}$$

$$E_3 = \frac{-13.6}{3^2} = -1.5 \text{ eV (the second excited state)}$$

$$E_4 = \frac{-13.6}{4^2} = -0.85 \text{ eV (the third excited state)}$$

These energy levels matched the experimental data to a high degree of accuracy.

The problems with Bohr's model

Despite its success regarding hydrogen, the Bohr model was limited in its application. We have not examined the detail of how he based his model on classical laws of orbiting systems and electrical forces. For these reasons, though, his model was only comprehensively applicable to one-electron atoms, these being hydrogen and ionised helium! It could not account for the spectrum of helium or any other elements. It modelled inner-shell electrons well but could not predict the higher energy orbits of multi-electron atoms. Nor could it explain the discovery of the continuous spectrum emitted by solids (now understood to be due to the interactions between their closely packed atoms). Soon even the emission spectrum of hydrogen was found to challenge Bohr's model. Some of the emission lines could be resolved into two very close spectral lines by using magnets. Bohr's theory could not explain this.

Although Bohr's model hasn't been able to quantify the exact electron energy levels of all atoms, its importance should not be underestimated. It signified a conceptual breakthrough. Bohr had the insight to correctly apply a quantum approach to the energy levels of atoms. He was the first to make links to the emerging areas of the quantum nature of electromagnetic radiation. Bohr had provided a conceptual framework on which future developments would be based.

Standing waves and the dual nature of matter

Around the same time that physicists were becoming aware that *photons* of electromagnetic radiation could be observed to behave very much like particles (particularly the high-energy X-rays utilised in the Compton effect), a French PhD student, Louis de Broglie, was approaching his area of research from a different perspective. De Broglie had a very strong belief in an underlying symmetry in nature. He reasoned that the dual nature of *light* must imply a dual nature for *matter*. He postulated that since light, once thought of as a continuous wave, was found to have a particle/photon nature, perhaps *particles of matter* might have some wave characteristics. Could he find evidence of the existence of *matter waves*? Although initially he was being purely speculative, de Broglie pursued his proposal with remarkable results:

De Broglie took Compton's relationship for the momentum of the *photon* and rearranged it as follows:

$$\lambda = \frac{h}{p}$$

Since he was dealing with particles of matter assumed to have mass m and speed v, the classical expression for momentum ($\boldsymbol{p} = mv$) was substituted into the above equation, giving:

$$\lambda = \frac{h}{p} = \frac{h}{mv}$$

He then took the important step of interpreting the equation, suggesting that λ is the *wavelength of the particle* itself. The term *matter wave* has since been employed. An effective way of picturing this is to say that the matter will behave *as if it had a particular wavelength value*. For example, an electron moving at a particular speed may be said to behave *as if* it had a wavelength of 1.50 nm. This electron would be expected to display some behaviour in common with other waves of this same wavelength.

The ***DE BROGLIE WAVELENGTH***, or wavelength of the matter wave, of a body of mass *m* moving with a speed *v* is given by:

$$\lambda = \frac{h}{p} = \frac{h}{mv}$$

where λ is the de Broglie wavelength of a particle (*m*), *h* is Planck's constant (6.63×10^{-34} J s), ***p*** is the momentum of the particle (kg m s^{-1}), *m* is the mass of the particle (kg), and ***v*** is the velocity of the particle (m s^{-1}).
Note that the electronvolt value of Planck's constant does not apply when using this formula!

De Broglie applied his approach to the discussion of Bohr's hydrogen atom. He viewed the electrons that were orbiting the hydrogen nucleus as matter waves. He suggested that the electron could only maintain a steady energy level is if it established a *standing wave*.

De Broglie's objective was to gain a picture of how the electron wave was arranged around the nucleus of the atom. He assumed that his standing wave must be three dimensional. To picture it, take a normal standing wave such as might be established in a vibrating string, and wrap it into a loop as shown in Figure 5.61.

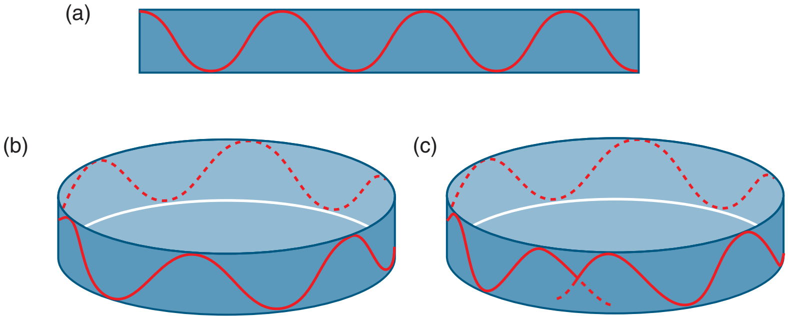


Figure 5.61

(a), (b) If a whole number of matter wavelengths 'fit' into the circumference of the electron orbit, then the wave reinforces itself, and a standing wave is produced. This can be interpreted by saying that an energy level may exist with this circumference. (c) Where an integral number of waves do not fit, destructive interference occurs, and the orbit cannot represent an energy level.

The only wavelength values that the electrons could 'have' were the wavelengths that fitted perfectly into the orbit:

Therefore the circumference of the orbit ($2\pi r$) must be equal to ($n\lambda$) where *n* is equal to 1, 2, 3, . . .

$$2\pi r = n\lambda \quad \text{and} \quad \lambda = \frac{h}{mv}$$

$$2\pi r = \frac{nh}{mv}$$

This was exactly the relationship that Bohr had assumed to exist and that led to the expression that we examined for the energy levels of the electrons of the hydrogen atom, namely $E_n = \frac{-13.6}{n^2}$.

Thus, by adopting a standing wave model and treating the orbiting electrons as matter waves, de Broglie had actually given a good reason for an assumption for which Bohr had been unable to provide a physical justification. By his use of a standing wave model for the quantised atom, de Broglie had provided further argument towards his belief in the dual nature of matter.

In conclusion. . .

Bohr's theory had been a mixture of classical and quantum theories, partially recognising the wave–particle duality of light and matter. However, it allowed new, more radical, theories to be developed by the physicists who followed him, such as Werner Heisenberg and Erwin Schroedinger. In our studies we have seen how early quantum ideas have explained the photoelectric effect and absorption and emission spectra. A comprehensive theory has since developed, explaining a wide range of phenomena. These examples only touch upon the very beginnings of the remarkable ideas of *quantum mechanics*.

Physics in action — Bohr and Heisenberg

In 1922, Danish Niels Bohr was 36 years old and had just been awarded the Nobel Prize for Physics for his work on the atom. As he gave a lecture, he was interrupted by a 20-year-old student, Werner Heisenberg, who had found a flaw in the maths of his argument! This meeting led to a strong friendship and working partnership between the two, culminating in Heisenberg being awarded his own Nobel Prize for Physics 10 years later in 1932. It was said that 'Bohr understood the world and Heisenberg knew the maths'. As a team their accomplishments were astounding.

In 1933 Hitler came to power in Germany. By 1939 it is reported that 29 of the 32 Nobel Prize winners who lived in Germany had fled. Heisenberg remained. Bohr lived in Copenhagen. Before the German invasion of Denmark, Bohr is credited with helping to provide an escape route for Jewish scientists. In September 1941 an infamous meeting took place. Heisenberg visited Bohr in the now occupied Denmark. Bohr later claimed that in this meeting Heisenberg confirmed that he was leading an atomic weapons program for the Germans, in a race to build a nuclear bomb. Bohr was shocked. The British government were sent their first confirmation of a German attempt to build an atom bomb. Bohr secretly fled occupied Denmark and joined the US nuclear weapons project in Los Alamos, New Mexico.

Heisenberg continued to work for the German government until the Russian invasion. Reportedly, German scientists had made enough progress for their seconded documents to provide a firm start to the Russian nuclear energy and arms program after the war. Many years later Heisenberg claimed that in the meeting with Bohr he was actually attempting to have a covert conversation with him, in order to establish an agreement that they should both undermine the work on nuclear weapons for humanitarian reasons. Heisenberg claimed that Bohr had misunderstood his double-speak. This issue was never resolved between the two former friends.

Figure 5.62
Although nuclear physics helped to bring World War II to a close, nuclear weapons testing continued.

5.5 SUMMARY Electromagnetic radiation and matter

- The energy of a photon is determined by its frequency:

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

- The energy absorbed or emitted by atoms provides clues about the atom's structure.
- The production of spectra suggests an internal structure to the atom. A line *emission* spectrum is produced by energised atoms, and an *absorption* spectrum is created when white light passes through a cold gas. The spectrum for an element is unique.
- Band emission spectra consist of many lines very close together in bands. They are characteristic of emission from excited molecules.
- Band absorption spectra consist of many lines very close together in bands. They are characteristic of coloured transparent solids and coloured solutions. Continuous emission spectra are emitted from hot solids and liquids. They contain a broad range of photon energies and are called black-body radiation. The distribution of photon energies and hence the overall appearance of colour depends on the temperature of the emitting substance.
- Fluorescence occurs when a substance absorbs high energy photons such as ultraviolet and re-emits the energy as visible light. Phosphorescence is similar to fluorescence but the excited state is metastable and the re-emission of energy takes place over a period of time.
- Niels Bohr suggested that electrons in atoms orbit the nucleus in specially defined energy levels, and no radiation is emitted or absorbed unless the electron can jump from its energy level to another. In this way, electron energies within the atom are quantised, since only certain values are allowed.
- An electron in an atom which jumps between energy levels m and n emits or absorbs a photon of energy:

$$E_{photon} = hf = E_m - E_n$$

Where the electron starts in level n and drops to level m, the photon will be emitted. Where the electron is promoted from the lower level m to the higher n, the photon energy will be absorbed.
- The Bohr model was limited in its application, it being only applicable to one-electron atoms, but it was a significant development because it took a quantum approach to the energy levels of atoms and incorporated the quantum nature of electromagnetic radiation.
- De Broglie viewed electrons as matter waves and his standing wave model for electron orbits provided a physical explanation for electrons only being able to occupy particular energy levels in atoms. He suggested that the only way that the electron could maintain a steady energy level was if it established a standing wave.
- Energy levels in the Bohr atom are analogous to the quantised modes of vibration (standing waves) that are known to occur in physical objects such as strings.

5.5 Questions

For the following questions use:

Planck's constant = 6.63×10^{-34} J s or 4.14×10^{-15} eV s

speed of light = 3.00×10^{8} m s^{-1}

charge on electron = 1.60×10^{-19} C

mass of electron = 9.11×10^{-31} kg

The following information applies to Question 1.

__________ ionisation level

__________ $n = 3$ second excited state

__________ $n = 2$ first excited state

__________ $n = 1$ ground state

1 When an electron makes a transition from $n = 3$ to $n = 1$, a photon of frequency 6.00×10^{14} Hz is emitted, while a transition from $n = 2$ to $n = 1$ results in the emission of a photon of frequency 4.00×10^{14} Hz.

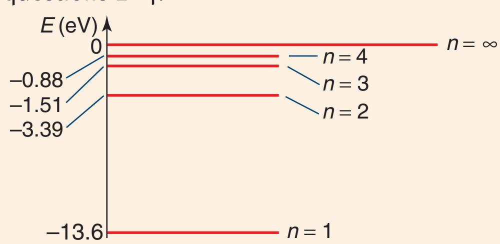
a What is the wavelength of the photon emitted in a transition from the $n = 3$ to $n = 2$?

b Which of the following best describes the light that corresponds to this wavelength?

A red **B** orange **C** blue

D infrared **E** X-rays

The following energy levels for atomic hydrogen apply to questions 2–4.



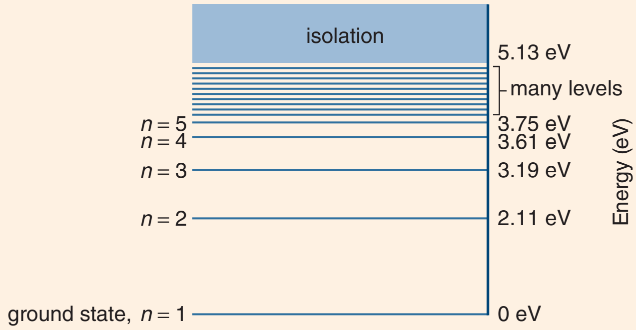
2 a Calculate the frequency of photons emitted when electrons in atomic hydrogen make the transition from the first excited state to the ground state.
b What is the wavelength of light that is emitted when the electron in atomic hydrogen goes from the second excited state to the ground state?
c What is the minimum energy (in eV) required to ionise a hydrogen atom from the ground state?

3 a If a hydrogen atom is initially in its ground state, determine the highest energy level to which it can be excited by collision with an electron of energy 12.5 eV.
b Would a photon of energy 12.5 eV be able to excite the hydrogen atom to the same level? Justify your answer.
c What would happen if a photon of energy 14.0 eV collided with the hydrogen atom while it was in the ground state?

4 A 12.8 eV electron collides with a hydrogen atom while in the ground state.
a What is the highest energy level that the atom could be excited to by such a collision?
b List all the possible photon energies that could be produced as the hydrogen atom returns to the ground state after such a collision.

5 Use Bohr's model of the atom to explain why all of the frequencies of light above the ionisation energy value for hydrogen are continuously absorbed.

The following information applies to questions 6–8. The diagram shows the energy levels of a sodium ion.



6 A sodium atom is in an $n = 3$ excited state. Calculate the shortest wavelength of light that this atom could emit.

7 Numerous excited sodium atoms have electrons in the $n = 5$ level. How many different photon energies may be observed as these atoms de-excite?

8 Are sodium atoms able to emit or absorb more frequencies of light? Explain.

9 The visible spectrum extends from approximately 700 to 400 nm. Are all the Balmer lines in the hydrogen spectrum visible?

10 What do de Broglie's matter wave concept and a bowed violin string have in common?

Figure 5.63
An X-ray photograph of a head with a nail embedded in it. The X-ray photograph enabled surgeons to accurately determine the position of the nail; it was then possible to remove the nail without causing permanent brain damage.

5.6 X-rays

An interesting region of the electromagnetic spectrum is known as X-radiation. This region is important because X-rays have many uses in medicine and industry and because the production of X-rays reveals some additional information about the structure of atoms.

The properties and uses of X-rays result from the fact that they are very penetrating electromagnetic radiation with characteristic high frequencies, small wavelengths and large photon energies. Their frequencies range from about 10^{17} Hz to 10^{19} Hz and lie in the electromagnetic spectrum between ultraviolet and gamma radiation. They are produced when high-energy electrons are decelerated rapidly and lose their kinetic energy as high-energy electromagnetic radiation. X-rays are also called bremsstrahlung, which is a German expression meaning 'braking radiation'.

Low-density substances are transparent to X-rays while more dense substances tend to attenuate X-rays by absorbing them. The greater the density of the material, the greater the attenuation or absorption. The higher the photon energy of the X-rays, the less a material will attenuate them. Consequently, soft tissue such as skin and muscle will readily allow X-rays to pass through them while bone does not readily allow X-rays to pass. By controlling the intensity and energy of incident X-radiation, physicians, dentists and veterinarians can make X-ray photographs of the interior of their patients including details of the internal structure of individual organs.

X-rays and low-energy gamma rays have enough energy to ionise atoms and molecules and this can cause damage to body tissues. The DNA molecule is very susceptible to damage and this can create problems if a cell with damaged DNA begins to divide. Most damage to DNA results in unviable cells which die. Sometimes, however, the cell can grow and divide and form a lump which does not have a specific function. Such growths can be malignant or benign. All cells are more sensitive to X-ray damage when they are dividing and this sensitivity is exploited when using X-rays to treat cancers. The very high cell division rate is the reason that bone marrow is more vulnerable than other tissues.

Protection from X-rays is necessary for anyone who might be frequently exposed. When physicians and dentists make X-ray photographs of their patients, they usually leave the room as it is usual for the switches that operate the X-ray equipment to be placed on the other side of a wall in another room. When the operator of X-ray equipment must be near the equipment, he or she will typically wear an apron lined with lead. A patient having a dental X-ray will usually have the rest of his or her body protected with a lead apron.

In industry, the penetrating ability of X-rays is used to examine metal joins, such as welds, for imperfections. The same techniques can be used to examine metal castings, forgings and assembled components for flaws. Even very slight differences in the materials will cause differences in the extent to which the X-rays are attenuated. This will reveal the variations or flaws in the materials if the X-rays are used to make a photograph or an image on a fluorescent screen.

X-ray crystallography has allowed research into the structure of quite complicated minerals and molecules. The intricate structure of proteins and molecules such as DNA has been revealed by X-ray diffraction patterns. Metal fatigue, the deterioration of metals under repeated stress, has been shown to be due to a gradual breaking down of the crystalline structure. X-ray diffraction techniques are possible because of the very small wavelength of X-rays. Interference patterns result from the spacing between atoms in a crystal or a molecule as these spacing are of similar orders of magnitude to the X-ray wavelengths.

Physics file

Wilhelm Conrad Roentgen discovered X-rays on 8 November 1895 and named them X-rays because he did not have any idea what kind of radiation they were. For the next seven weeks he lived, ate and slept in his laboratory while he thoroughly investigated the properties of X-rays. He invented X-ray photography and his wife's hand was the first object ever to be photographed in this way. The image of her own skeleton frightened her as she thought that it was a premonition of her own death. The first announcement of his discovery was a short paper given to the Wurzburg Physical Medical Society on 28 December 1895. Great excitement and interest followed and in 1896 over 1000 articles relating to X-rays were published. Roentgen refused to patent, or develop for commercial gain, any of his discoveries, insisting that they should be for the benefit of all. He was the first person to be awarded the Nobel Prize in Physics.

Figure 5.64
The world's first X-ray photograph shows Roentgen's wife's hand wearing a ring.

X-ray production

An X-ray tube uses a high voltage to accelerate electrons before they are allowed to strike a metal target and lose their kinetic energy. This results in a continuous spectrum of X-ray energies with features resulting from both the size of the accelerating voltage and the particular metal used for the target. Figure 5.65 shows the X-ray spectrum of an X-ray tube operating with an accelerating potential of 50 kV and a target made from molybdenum. The peaks labelled K_α and K_β are characteristic of the molybdenum target and are actually a line spectrum superimposed over the continuous spectrum. The minimum wavelength, labelled λ_{min}, is a consequence of the 50 kV accelerating voltage applied across the X-ray tube.

In order to understand the operation of the X-ray tube and the features of the spectrum, the acceleration of the electrons by the large voltage across the tube must first be considered. Then what happens to the electrons when they strike the target can be considered.

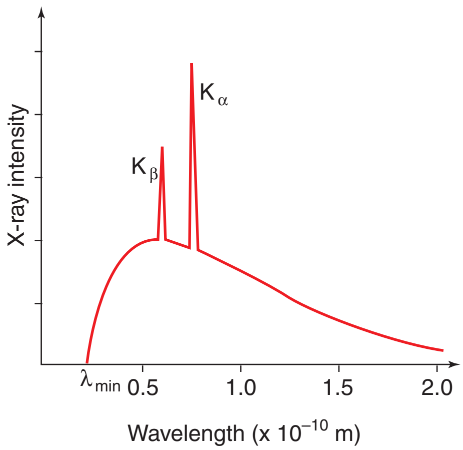


Figure 5.65
The spectrum obtained from a 50 kV X-ray tube with a molybdenum target anode.

Figure 5.66 is a schematic diagram of an X-ray tube. The heater voltage heats the cathode and the ejected electrons accelerate across to the target anode. The large voltage between the cathode and the anode creates an intense electric field which accelerates the electrons until they strike the target anode. The electrons striking the target have a range of energies and the maximum kinetic energy achieved by the electrons is equal to the product of the accelerating voltage and the charge on an electron.

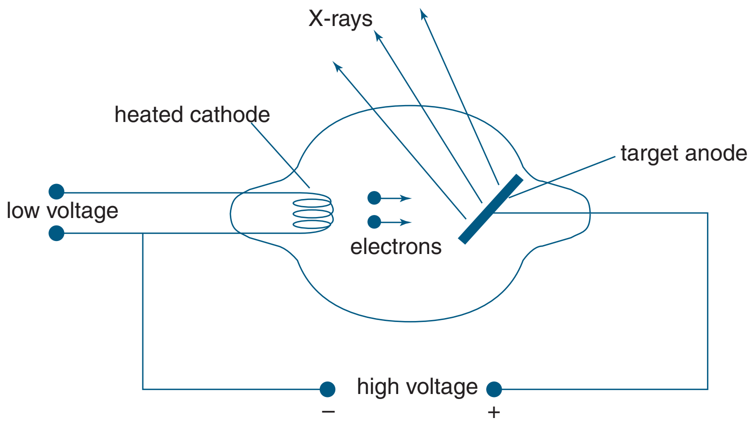


Figure 5.66
A schematic diagram of an X-ray tube. The electrons are accelerated by the high voltage before releasing their kinetic energy as X-radiation when they collide with the target anode.

When the electrons strike the target they can lose their energy in a number of ways. If an electron collides with the target and stops suddenly, it loses all of its kinetic energy at once and emits a single photon equal to the kinetic energy of the electron. Hence, the most energetic photon that can be obtained from an X-ray tube has a kinetic energy equal to the most energetic electrons accelerated by the accelerating voltage. Consequently the greatest possible photon energy is also the product of the accelerating voltage and the charge on an electron.

The ***LARGEST PHOTON ENERGY*** produced by an X-ray tube is given by:

$$E_{ph(max)} = e\Delta V$$

where $E_{ph(max)}$ is the energy, in joules (J), of the most energetic photons emitted from the X-ray tube, e is the charge on an electron in coulombs (C) and ΔV is the accelerating potential difference of the X-ray tube.

When electrons strike the target they will typically slow down in a number of collisions with individual atoms and lose their energy in more than one step. Consequently they will create a number of photons each with less energy than the maximum possible photon energy. This explains why X-ray spectra are continuous and have a wide range of photon energies up to a maximum value. The interactions that X-rays exhibit with matter will be explored later in this section.

As is the case for photons in other parts of the electromagnetic spectrum, the energy of an X-ray photon is the product of Planck's constant and the frequency of the radiation.

Physics file

The ***MAXIMUM KINETIC ENERGY*** of an electron striking the target anode of an X-ray tube is given by:

$$E_{k(max)} = e\Delta V$$

where $E_{k(max)}$ is the maximum kinetic energy (J), ΔV is the accelerating potential difference in volts (V) and e is the charge on an electron in coulombs (-1.60×10^{-19} C).
Note that the negative charge on the electron is cancelled by the negative accelerating potential difference as the electric field is in the opposite direction to the acceleration of the electron.
$E_{k(max)}$ can also be expressed in electron volts, V is the accelerating voltage in volts and e has the value of 1.

The ***MAXIMUM PHOTON ENERGY***, maximum frequency and minimum wavelength for the X-ray spectrum are given by:

$$E_{ph\,(max)} = hf_{max} = \frac{hc}{\lambda_{min}} = e\Delta V$$

where $E_{ph(max)}$ is the maximum X-ray photon energy (J), h is Planck's constant (6.63×10^{-34} J s), f_{max} is the maximum frequency of the radiation in hertz (s^{-1}), c is the speed of light in a vacuum (3.00×10^{8} m s^{-1}), λ_{min} is the shortest wavelength of the radiation (m), V is the accelerating potential difference of the X-ray tube (V) and e is the charge on an electron (-1.60×10^{-19} C).

Traditionally, graphs of spectral intensity have wavelengths on the horizontal axis. Since wavelength is inversely proportional to photon energy, the spectrum has a minimum photon wavelength corresponding to the maximum photon energy.

Characteristic peaks in X-ray spectra

As detailed above, an X-ray spectrum has a line spectrum superimposed over the continuous spectrum. These peaks of intensity are characteristic of the target anode material. It is possible for the incoming electrons to strike an atom with enough energy to remove an electron from an inner orbital. The energy levels or electron orbitals are traditionally labelled with letters. The first level is K, the second level is L and so on. If, for example, an electron is removed from the K shell of an atom of the target then this creates a highly unstable situation and another electron from a higher energy shell such as the M or N shell or from outside the atom will immediately fall into the K shell. This is a similar process to the one described in section 5.5 where an electron from a higher orbital falls back to the 'valence' orbital to emit a photon of visible or ultraviolet light. In this case, however, the energy transition down to the K shell is so large that the emitted photon is an X-ray. Figure 5.68 shows the transitions which result in the K_α and the K_β peaks in the spectrum of molybdenum. The sizes of the energy transitions between energy levels or shells vary for different elements. Consequently, the peaks in the X-ray spectrum will be in unique positions for each element.

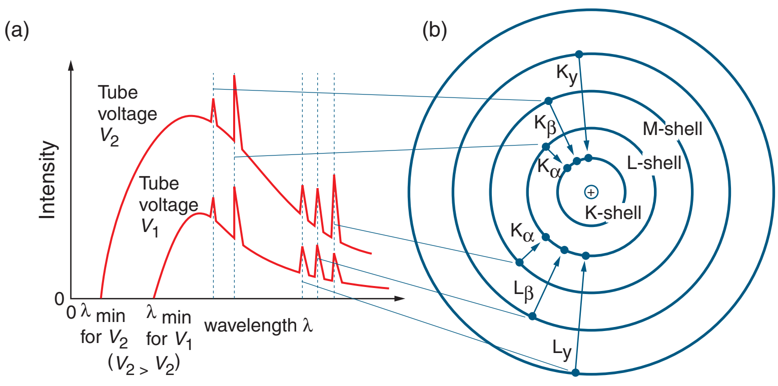


Figure 5.68

The spectrum produced by an X-ray tube depends on the accelerating voltage and the element in the target. λ_{min} depends on the maximum energy of the bombarding electrons. The peaks in the spectra are a line spectrum produced when inner orbital electrons are displaced and replaced by electrons from the higher energy shells.

Physics file

Only about 1% of the energy in the electron beam of an X-ray tube is converted to X-radiation. The other 99% is converted to heat. Consequently, design features such as a rotating target anode or an oil-cooled target anode must be built into X-ray tubes. The oil-cooled version works a little like the water cooling system of a car engine, as oil is circulated through channels in the target anode. A rotating target anode avoids overheating of the target by making sure that the electron beam does not strike one place of the target anode for too long.

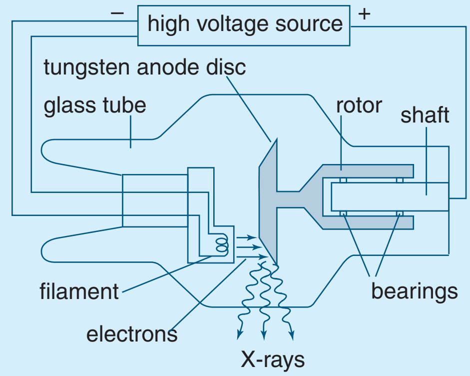


Figure 5.67

An X-ray tube typically has a tungsten anode which rotates so that the electrons do not continuously heat, and therefore melt, the same section of metal.

Practical activity

33 Diagnostic X-rays

✓ Worked Example 5.6A

Figure 5.65 shows the spectrum of X-rays that would be obtained from a 50.0 kV X-ray tube with a molybdenum target.

a Calculate the maximum kinetic energy of the bombarding electrons in both electronvolts and joules.
b Calculate the maximum energy of the emitted photons in joules.
c Calculate the maximum speed of the bombarding electrons.
d Calculate the maximum frequency of the emitted X-rays.
e Calculate the minimum wavelength of the emitted photons.
f Calculate the photon energy of the K_α emission in joules and electronvolts.

Solution

a The accelerating voltage is 5.00×10^4 V so the maximum possible energy of the accelerated electrons is 5.00×10^4 eV. In joules this is:

$$E_{k(max)} = (5.00\times10^{4})\times(1.60\times10^{-19})$$
$$= 8.00\times10^{-15}\ \text{J}$$

since 1 eV = 1.6×10^{-19} J.

b The maximum energy of the emitted photons is the same as the maximum kinetic energy of the bombarding electrons:

$$E_{ph(max)} = E_{k(max)}$$
$$= 8.00\times10^{-15}\ \text{J}$$

c The maximum speed of the bombarding electrons is found from the maximum kinetic energy:

$m_e = 9.11\times10^{-31}$ kg

$E_{k(max)} = 8.00\times10^{-15}$ J

$$E_{k(max)} = \tfrac{1}{2}mv^2_{max}$$
$$v_{max} = \sqrt{\frac{2E_{k(max)}}{m}} = \sqrt{\frac{2(8.00\times10^{-15})}{9.11\times10^{-31}}}$$
$$= 1.33\times10^{8}\ \text{m s}^{-1}$$

d The maximum frequency of the X-rays is given by:

$h = 6.63\times10^{-34}$ J s

$E_{ph(max)} = 8.00\times10^{-15}$ J

$$E_{ph(max)} = hf_{max}$$
$$f_{max} = \frac{E_{ph(max)}}{h} = \frac{8.00\times10^{-15}}{6.63\times10^{-34}}$$
$$= 1.21\times10^{19}\ \text{Hz}$$

e The wavelength can be found from frequency and the speed of light using the wave equation:

$f_{max} = 1.21\times10^{19}$ Hz

$c = 3.00\times10^{8}$ m s^{-1}

$$\lambda_{min} = \frac{c}{f_{max}} = \frac{3.00\times10^{8}}{1.21\times10^{19}}$$
$$= 2.49\times10^{-11}\ \text{m}$$

f The K_α line appears to correspond to a wavelength of 0.75×10^{-10} m

$\lambda_\alpha = 0.75\times10^{-10}$ m

$c = 3.00\times10^{8}$ m s^{-1}

$h = 6.63\times10^{-34}$ J s

$$E_\alpha = \frac{hc}{\lambda_\alpha} = \frac{(6.63\times10^{-34})(3.00\times10^{8})}{0.75\times10^{10}}$$
$$= 2.65\times10^{-15}\ \text{J}$$
$$= 1.66\times10^{4}\ \text{eV}$$

Synchrotron radiation

What is synchrotron light?

Synchrotron light is the name given to electromagnetic radiation emitted when charged particles, such as electrons, are accelerated in curved paths. For high-energy electrons, the photons emitted have energies ranging from the infrared through to soft and hard X-rays. Synchrotron-generated light has a number of advantages that make it suitable for a range of experimental techniques.

Synchrotron radiation has the following characteristics:

- high intensity or brightness—hundreds of times brighter than from standard X-ray tubes
- broad spectral range—ranging from infrared light to hard X-rays
- a high degree of collimation—having low beam divergence.

It is:

- tunable—meaning that required frequencies can be selected
- pulsed—emitted in very short pulses of less than a nanosecond
- highly polarised—either linearly, circularly or elliptically polarised.

Physics file

In 2003, space physicist Joseph Dwyer at the Florida Institute of Technology reported that a mysterious property of lightning had been confirmed. Just before a lightning strike, a burst of X-rays can usually be detected. The X-ray burst lasts for around one ten-thousandth of a second and the photons emitted have energy of around 10 000 eV. This research may mean that physicists need to rethink how lightning is made.

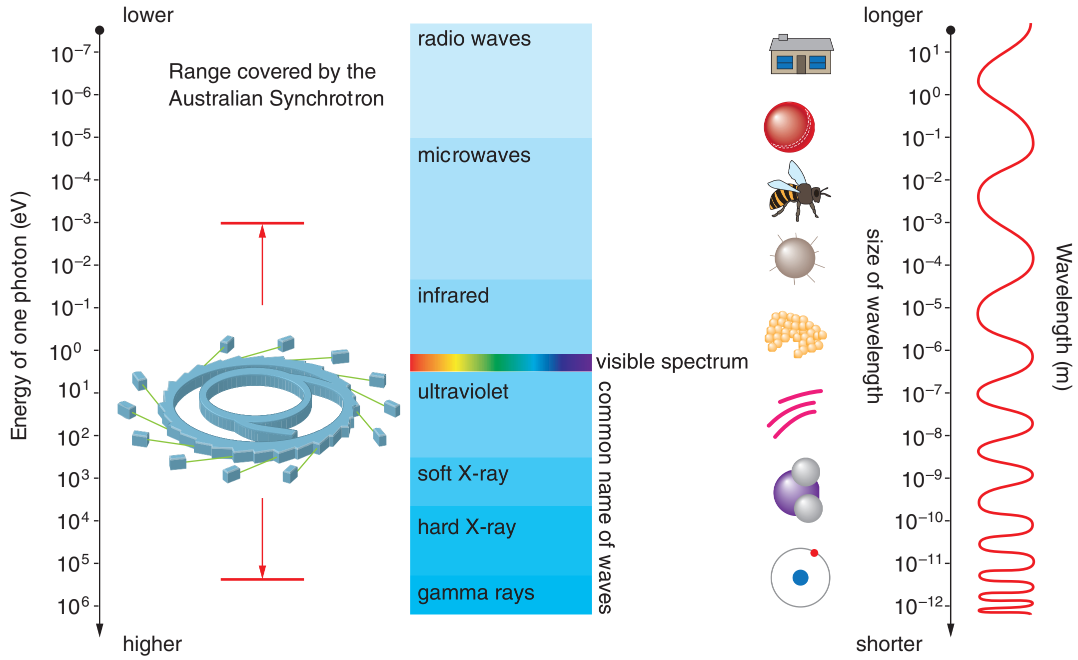


Figure 5.69
The range of wavelengths and photon energies produced as synchrotron radiation.

The specific attributes of synchrotron radiation make it useful in a wide range of techniques. The range of photon energies produced in the Australian Synchrotron can be seen in Figure 5.69. Such electromagnetic radiation has wavelengths corresponding to the dimensions of cells, viruses, proteins and atoms. From Figure 5.69 it can be seen that visible light has a wavelength much larger than the size of proteins. Shorter wavelengths enable scientists to explore the structure of similarly small objects. An ordinary light microscope is incapable of resolving such structures, due to the much longer wavelength of visible light. The short-wavelength X-rays produced in synchrotron light are an ideal tool for examining structures at a cellular or atomic level. High-energy X-rays can resolve down to scales of 10^{-10} m, or 1 Å (angstrom), the size of individual atoms. The brightness and monochromatic nature of synchrotron radiation make it ideal to delve into the make-up of crystalline structures, using a technique called X-ray diffraction.

The high intensity of synchrotron light is useful in experimental situations because it means a particular analysis can be completed in a far shorter time than would be the case if electromagnetic radiation from another source, such as an X-ray tube, was used. Some experiments can also be conducted *in situ*, or using a sample in its natural state, rather than after some treatment. In this way, it is possible to use synchrotron light to study how some processes change in real time.

Another advantage of synchrotron light is that a continuum of radiation is produced. Different elements will absorb energy of a specific frequency. By being able to select particular wavelengths of synchrotron light, researchers can select the best wavelength or range of wavelengths for a specific technique or analysis.

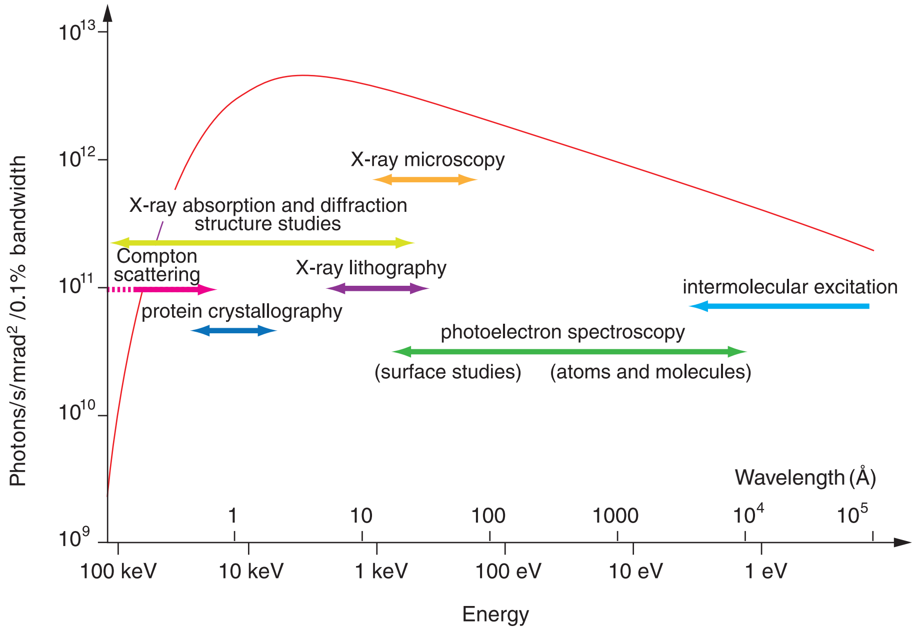


Figure 5.70

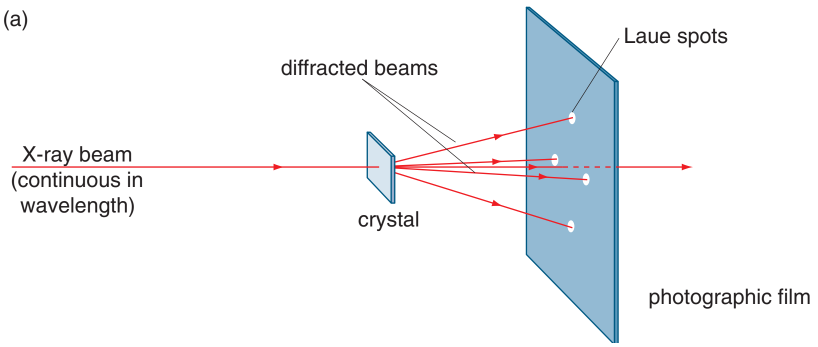
Almost all areas of science have benefited from the use of synchrotron radiation. The range of techniques possible using various bands of synchrotron radiation are highlighted in this figure. This graph illustrates the range of synchrotron radiation emitted from the Synchrotron Radiation Source at the Daresbury Laboratory, UK. Specific applications of bands of radiation are marked.

A range of techniques is being pioneered in a field called phase-contrast X-ray imaging. Soft tissue, such as tumours, cartilage, ligaments and skin, does not provide much detail when X-rayed in a conventional sense. Phase-contrast X-ray imaging techniques utilise synchrotron X-rays to produce more contrasting and hence more detailed images of such soft tissue body parts.

The linearly polarised synchrotron beam can be converted into a circularly polarised beam and utilised in Compton scattering experiments on magnetic samples. The energy and highly collimated nature of the beam make it suitable for a wide range of diffraction pattern analyses through which computers can be used to generate three-dimensional models of protein macromolecules. Scientists are also developing high spatial resolution imaging techniques that enable synchrotron light to be used in the diagnosis of tumours. Microscopic X-ray lithography is another use of synchrotron light. High-resolution X-ray spectroscopy experiments, studies of small crystals and X-ray imaging began in the 1970s. These rely on the use of X-rays of varying energies. Before exploring some techniques of analysis, we will have a closer look at X-rays themselves.

Properties of X-rays

In 1912, the German scientist Max von Laue performed an experiment that helped to define the nature of X-rays. He knew that they were not charged particles because they were not deflected by electric or magnetic fields. Other scientists had failed in their attempts to study whether X-rays could be diffracted by using gratings similar to those used for light experiments. Von Laue realised that the reason this was not successful could be because X-rays had wavelengths much smaller than those of visible light. Because of this, no noticeable diffraction would be observed. He suggested that the spacing of atoms in a crystal such as sodium chloride could be of the same order of size as the wavelength of X-rays. Two of von Laue's colleagues, Knipping and Friedrich, tested his idea using the experimental setup shown in Figure 5.71a.



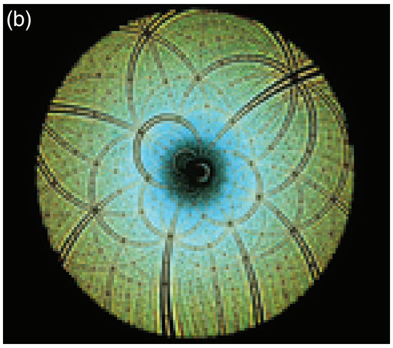


Figure 5.71

(a) The arrangement of atoms in the crystal can be determined from an analysis of the Laue spots produced in the diffraction pattern. (b) Laue X-ray diffraction pattern from a crystal of the enzyme rubisco (ribulose biphosphate carboxylase oxygenase) from a plant.

It was shown that the thin beam of X-rays was diffracted by the crystal. After a long exposure, a characteristic diffraction pattern, known as Laue spots, was found on the photographic plate (Figure 5.71b). This verified that X-rays are wavelike in nature, and have wavelengths of about 10^{-10} m (or 1 Å). They were later shown to be a type of electromagnetic radiation.

The general properties of X-rays can be summarised as follows.

- They have wavelengths similar to atomic spacings.
- They travel in straight lines.
- They readily penetrate matter (but less so in higher density materials and elements of higher atomic number).
- They are not deflected by electric or magnetic fields.
- They cause electrons to eject from a material through the photoelectric effect.

When is an X-ray not an X-ray?

It is likely that you are most familiar with the use of X-rays to detect a broken bone or hidden dental problems. You may then wonder why we would use a large-scale facility such as a synchrotron to create something that can be done in a dentist's consulting room.

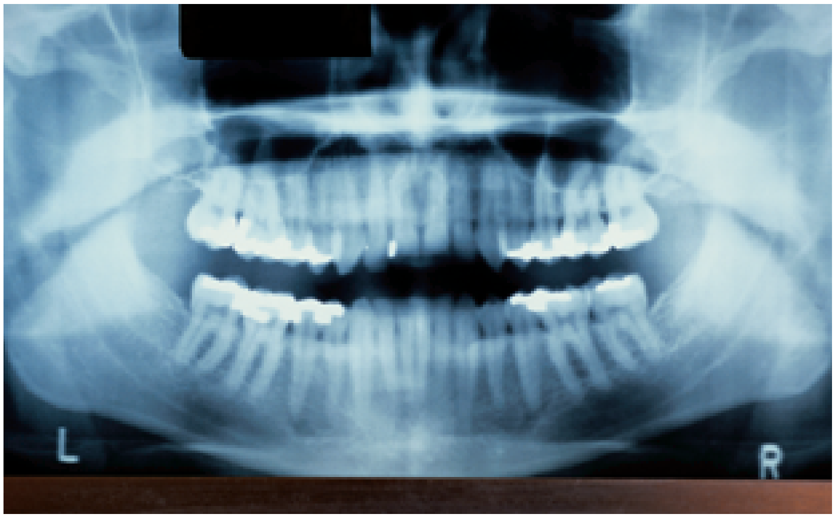


Figure 5.72
Say cheese! The toothy grin of a dental X-ray.

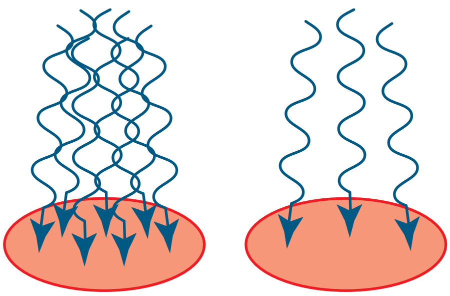


Figure 5.73
The X-rays delivered from a synchrotron source are far brighter than X-rays produced by an X-ray tube because a higher intensity is concentrated in a smaller area.

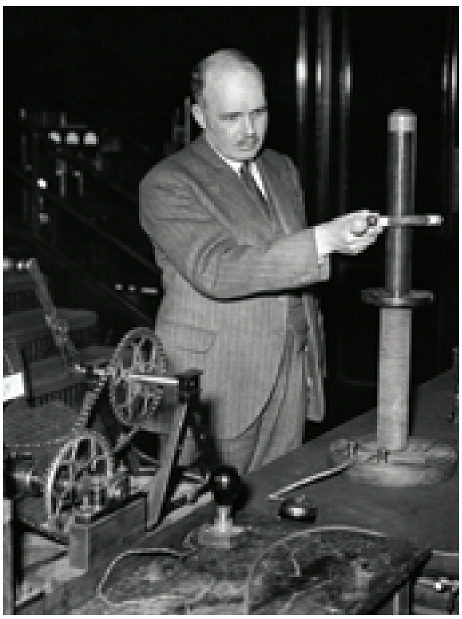


Figure 5.74
Sir Lawrence Bragg (1890–1971)—known as the father of the science of X-ray crystallography.

The types of X-rays that are produced in an X-ray tube are very useful in their diagnostic roles. The characteristics of the spectrum of wavelengths of X-rays produced in a synchrotron differ from those produced in an X-ray tube. A conventional X-ray is low in intensity and not coherent, having a greater spread of frequencies. Synchrotron X-rays are collimated, coherent and have higher intensity. In addition, any desired wavelength within the emittance spectrum can be selected by using a monochromator crystal to tune out particular synchrotron X-rays, whereas conventional X-ray sources produce radiation at only a few specific wavelengths. Synchrotron X-rays have suitable energies to interact with many common smaller atoms, like carbon and oxygen, whereas conventional X-rays have specific energies that will interact with heavier elements. Synchrotron X-rays are also around 100 million times brighter than those from conventional sources. The high brightness is due to the synchrotron X-rays being concentrated in a much smaller area, as shown in Figure 5.73. We can also say that a synchrotron beam has greater brilliance than a traditional X-ray beam, meaning that it has a higher intensity per unit area.

X-ray diffraction

Sir Lawrence Bragg was born on 31 March 1890 in Adelaide. He was the son of Sir William Bragg, a professor of mathematics and physics. After completing an honours degree in mathematics at the University of Adelaide, Sir Lawrence Bragg studied physics at Trinity College, Cambridge. It was here that he formulated what we now call Bragg's law. This law forms the basis of the technique of X-ray crystallography, through which we can determine the structure of complex molecules. For their services in the analysis of crystal structure by means of X-rays, Sir William and Sir Lawrence Bragg were jointly awarded the Nobel Prize in Physics in 1915.

The father and son team developed a technique for analysing diffraction patterns. If a collimated X-ray beam falls on a single plane of atoms in a crystal such as sodium chloride, then each atom of the layer will scatter a small portion of the beam. The beam is scattered in many directions. For most directions, the scattered wavelets reinforce destructively. When the angle of incidence of the beam onto the crystal atom equals the angle of reflection (or when the path difference between incident and reflected beams is equivalent to a multiple of beam wavelengths), then the wavelets reinforce constructively. For these particular angles, a reflected beam can be detected.

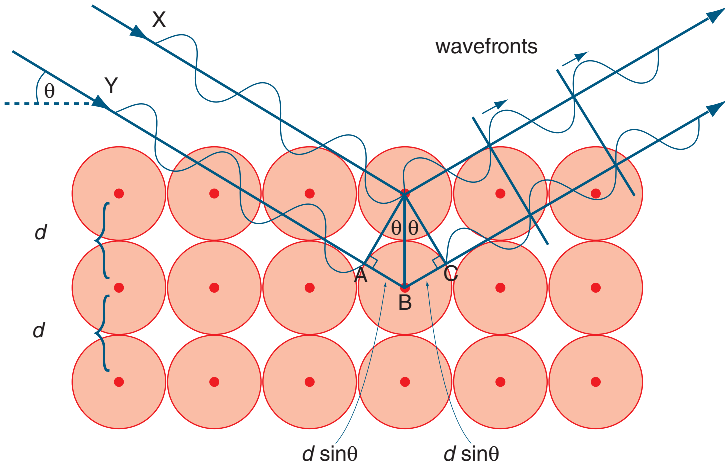


Figure 5.75
Wavefronts X and Y are scattered from layers of atoms of separation, *d*.

For example, consider the case of two X-ray waves, X and Y, entering a layer of atoms in a crystal as shown in Figure 5.75. In order for significant diffraction to occur, assume that X and Y have a similar wavelength to the distance between layers of atoms, *d*. The waves X and Y are initially in phase and are reflected by the atoms of the crystal. The rays will interfere constructively, producing a high-intensity reflected beam when they are scattered in phase. This occurs when the extra distance travelled by wave Y is equivalent to a whole number of wavelengths of the X-ray beam; that is, if $AB + BC = n\lambda$, where n is an integer and λ is the wavelength of the X-ray beam.

By using X-rays of a known wavelength, and rotating the crystal source through all angles with respect to the incoming beam, a pattern of intensity peaks may be produced, built up from rays reflecting from many layers in the crystal. The peaks produced can be examined, with the smallest glancing angle for which a maximum is recorded corresponding to $n = 1$.The complete diffraction pattern generated in such a process is unique to a particular crystal. This information can be used to determine the spacing between atoms in the crystal structure.

Sir Lawrence Bragg used the X-ray spectrometer designed by his father to analyse many crystals. Figure 5.76 shows the experimental set-up used. X-rays pass through two slits that collimate, or narrow, the beam. They reflect off a sample crystal and then pass through a third slit and into a detector where an ionisation current is registered. The magnitude of this current provides a measure of the intensity of X-rays reflected. The crystal and detector are rotated, so that the angle of rotation of the detector is twice the angle of rotation of the crystal. This set-up keeps the angles of incidence and reflection equal. Variations in the ionisation current indicate the positions of the glancing angles for which Bragg's law is satisfied, i.e. $2d\sin\theta = n\lambda$. Knowing the wavelength of the X-rays, we can calculate d once the angles are measured. Alternatively, if d is known, then the wavelength of X-rays can be calculated.

Physics file

In the course of this discussion on X-ray diffraction, the scattered beam has been referred to as being partially reflected by the target material. This is because the scattered portion is detected at an angle of reflection equal to the angle of incidence, just as in reflection from a plane mirror. In actual fact, the scattering atoms absorb and then re-emit this fraction of the incident X-ray beam.

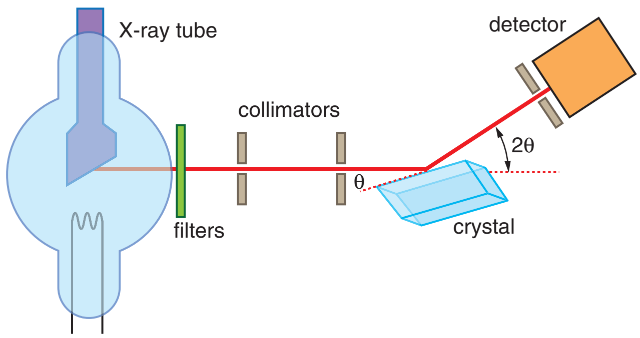


Figure 5.76
The Bragg spectrometer operates so that as the crystal rotates, the detector rotates through twice the angle. This set-up means that the detector remains in a position to register Bragg diffraction.

Interactions between X-rays and matter

Thomson scattering

We have studied the production of synchrotron light, particularly focusing on X-rays. In any collision, momentum and the total energy of a system are conserved. In the special case of kinetic energy being conserved, we say that the collision is elastic. Elastic collisions are not really observed in everyday life, because small amounts of kinetic energy are usually lost to heat or sound. Collisions between objects with little friction, like billiard balls, are almost elastic. *Bragg diffraction* of X-rays off a layer of atoms is completely elastic. The X-rays scatter from the atoms with no loss of kinetic energy; they do not lose any energy to the scattering atoms. Elastic scattering of this type is known as *Thomson scattering*. It occurs only with X-rays of relatively low energies of some 100 keV or less.

It is also possible for X-rays to undergo interactions with matter and lose some of their energy to the scattering atoms. These collisions are inelastic, because the X-rays interact with the matter with which they collide.

The photoelectric effect

Electrons are ejected from a metal when hit by light of a sufficiently high frequency. This is called the *photoelectric effect* and was studied earlier -in this chapter. In this situation, X-ray photons interact with bound electrons in the metal. Each electron can only interact with a single photon. It absorbs either all or none of the photon energy. If it absorbs all, then a photoelectron is ejected. The term *photoelectron* is used to acknowledge that light was responsible for the release of the electron.

Photoelectron flow is measured by an experimental set-up similar to that shown in Figure 5.77.

Experiments dating from 1887 confirmed that photoelectrons were only emitted once the incident light was above a minimum threshold frequency, f_0. Einstein received the Nobel Prize in Physics in 1921 for his explanation of this effect. The photoelectric equation links his ideas together.

If the voltage, V, of the circuit is reversed, the photocurrent does not immediately fall to zero. If this reversed potential difference is increased, then eventually photoelectrons cease to be ejected. The value beyond which this occurs is called the stopping voltage, V_0. The maximum kinetic energy of an ejected photoelectron can be shown to be:

$$E_{k(max)} = eV_0$$

This value is therefore independent of the intensity of incident light and depends only on its frequency.

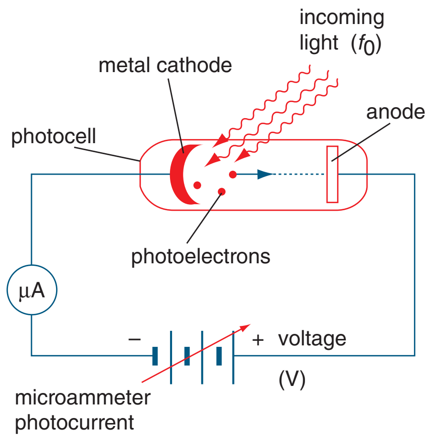


Figure 5.77
Incident light hitting the metal cathode will result in the emission of photoelectrons, provided that the frequency of the incoming light is sufficiently high. The voltage, *V*, can be varied continuously and also reversed by a switching arrangement.

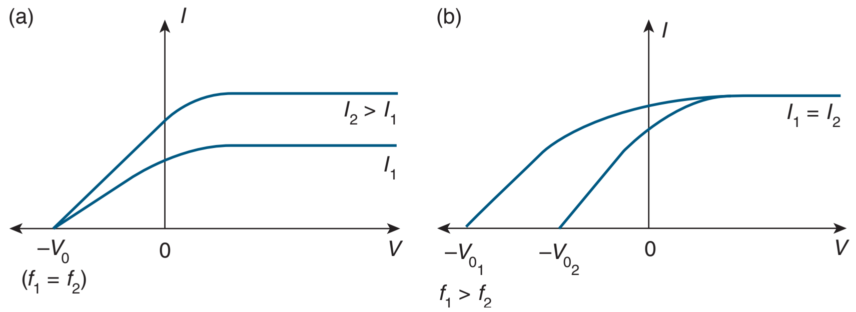


Figure 5.78

Photoelectric current plotted as a function of the applied voltage between the cathode and the anode. A standard monochromatic light source where $f > f_0$ shows that with a forward potential, every available photoelectron is included in the current. With a reverse potential, the number of photoelectrons decreases until none is collected at the stopping voltage, $-V_0$. (a) For brighter light of the same frequency, there is a higher photoelectric current, but the same stopping voltage. (b) For light with a higher frequency, there is an increase in the stopping voltage.

Physics file

The process of pair production describes the creation of an electron–positron pair. For pair production to take place, an incoming photon must exceed a minimum threshold energy value, equivalent to at least the mass of two electrons. Because one electron has a rest mass energy equivalent to 0.51 MeV (according to Einstein's $E = mc^2$), it follows that to produce two electrons, the photon must have an energy of at least 1.02 MeV. Usually, the positron and electron mutually annihilate into two 511 keV photons, which travel in opposite directions. Pair production is a direct conversion of radiation into matter and is the major means by which gamma rays are absorbed into matter.

The photoelectrons ejected may undergo further interactions with nearby atoms. This can create further ionisations. As holes left by ejected electrons are filled by less strongly bound outer-shell electrons, a wave of emission of fluorescence X-rays can result. These are distinctive of the absorbing atom. This process is used with synchrotron sources to examine samples for trace elements and is important for non-destructive investigation of archaeological and geological samples.

Compton scattering

Bragg scattering of X-rays involves no transfer of energy to the electrons that scatter them. In contrast, the photoelectric effect consists of a complete transfer of energy from an X-ray photon to an electron. Compton scattering describes an X-ray–electron collision in which some of the incoming photon energy is transferred to the electron. These collisions are inelastic. The discovery of this process gave further evidence to the fact that photons possess momentum, and can behave in a particle-like manner.

US physicist Arthur Holly Compton (1892–1962) discovered X-ray scattering of this type by using an apparatus similar to that shown in Figure 5.79.

Compton found that X-rays hitting a graphite target either emerged unchanged and unscattered or were scattered and emerged with less energy than they initially possessed. These scattered X-ray photons also emerged with a longer wavelength, corresponding to this loss of energy. The energy lost was transferred in the collision to an electron, which was ejected from the graphite. This occurred so that energy and momentum were conserved. Realising this, Compton then investigated the momentum of photons. He and British physicist Charles Wilson shared the 1927 Nobel Prize in Physics for their work in this area.

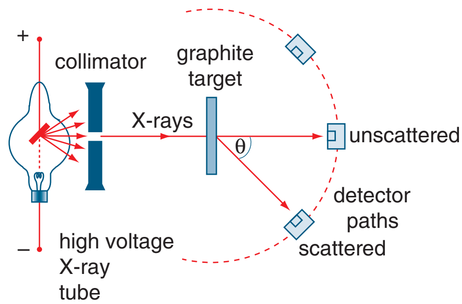


Figure 5.79

X-rays were made to collide with a graphite target. Some passed through unscattered and unaltered. Others were scattered and emerged with less energy than the unscattered X-rays.

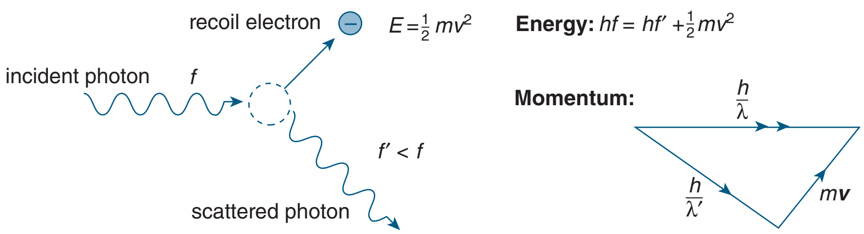


Figure 5.80

In the case of Compton scattering, energy and momentum are given up by some incident X-ray photons to electrons that are then ejected. These collisions are inelastic.

Physics in action — X-ray imaging

The first X-ray photographs were made with black and white photographic film identical to the film used in cameras that make photographs with visible light. Unfortunately, photographic film is not very sensitive to X-radiation and long exposures to dangerous levels of X-rays are required. The first such photograph, made in 1895, of Mrs Roentgen's hand was made with a 15-minute exposure! Modern X-ray film is coated on both sides with light-sensitive chemicals, and intensifying screens are used to increase the effectiveness of the film. The intensifying screens fluoresce when exposed to X-rays. In this way, the image is actually recorded with visible light, to which the photographic film is very sensitive. This arrangement, as shown in Figure 5.81, requires only short exposures to quite low levels of X-rays to make a photograph of a patient's bones or teeth. This type of X-ray imaging is essentially shadowing and reveals only two-dimensional information about the patient.

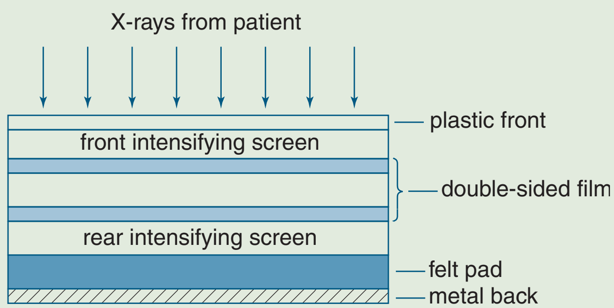


Figure 5.81
Intensifying screens fluoresce and provide visible light exposure for the X-ray film. Two light-sensitive layers also make the film effectively more sensitive to the X-rays.

Computer assisted tomography (CAT) Scanning involves sending a narrow beam of X-radiation through the body to a detector which records the transmitted intensity. Measurements are made at close intervals as the entire apparatus is moved sideways across the body. The entire arrangement is then rotated about 1° and another complete scan across the body is recorded. This process is repeated until the whole apparatus has been turned through 180°. The intensity of the beam for each position of each scan and for each angle of scanning is sent to a computer. The computer is programmed to use this information to construct an image of a 'slice' of the body. The image is then displayed on a TV monitor.

Figure 5.82
The CAT scanner penetrates the body with a narrow X-ray beam. The absorption information from many different angles is fed to a computer which processes it to create images of cross-sections of the body.

X-rays can have some damaging physiological effects as well as providing great benefits for people. There are both risks and benefits associated with the use of many regions of the electromagnetic spectrum. Table 5.2 outlines some of these.

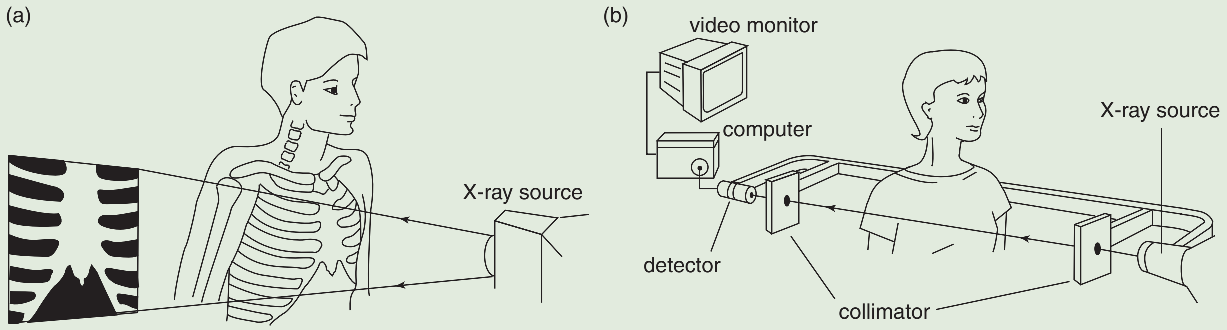


Figure 5.83

(a) A simple X-ray photograph makes a silhouette of dense parts of the body such as bones. (b) The arrangement of equipment for a CAT scan creates the information needed for a computer to build up three-dimensional information about the body.

table 5.2 Benefits and risks associated with electromagnetic radiation

Region of spectrum	Uses and benefits	Physiological risks
50 Hz radiation from electricity transmission lines	Countless obvious examples of benefits of electricity being widely available to the general population	No known or proven physiological effects of this frequency. Some statistical research in various parts of the world has tentatively linked this radiation with leukemia.
Radio waves from TV and radio transmitting stations	Mass communication	Some statistical evidence that very high intensity radiation in the higher frequency part of the radio spectrum may cause cancers.
Microwaves from radar transmitters, communications networks and ovens	Mass communication Fast cooking	Microwaves cause atoms to vibrate and consequently get hotter. This can be dangerous in living tissue. There is some evidence that exposure to microwaves can cause cataracts.
Infrared radiation from hot objects	Obviously useful in many situations where an increase in temperature is required by radiant heat.	Large doses can cause severe burns.
Visible light from the Sun and artificial sources of light	Vision and photosynthesis	No known physiological risks for people.
Ultraviolet radiation from the Sun and from UV lamps	Can be used to kill bacteria and to treat some skin conditions	More energetic ultraviolet can be ionising. There is significant evidence that this radiation can cause skin cancer.
X-rays from X-ray tubes	Medical imaging, airport security etc.	X-radiation is very ionising and can cause serious damage to cells including changes to DNA which in turn can lead to cancer.
Gamma radiation from radioactive materials or from equipment similar to X-ray tubes	Low-energy gamma rays are used for similar purposes to X-rays.	Low-energy gamma rays have all of the same dangers as X-rays. High-energy gamma rays can affect the nuclei of atoms.

5.6 SUMMARY X-rays

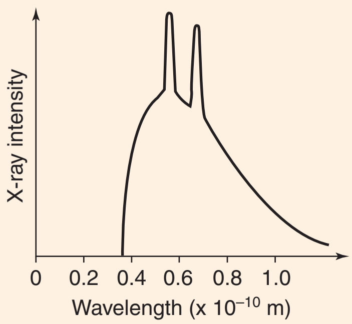
- X-rays are high photon energy, high-frequency, short-wavelength electromagnetic radiation.
- X-rays are produced when high-energy electrons in an X-ray tube collide with a target and lose their kinetic energy. The electrons are first accelerated by a large voltage and then hit a target anode. The maximum energy of the emitted photons depends on the size of the accelerating voltage and is equal to the product of the accelerating voltage and the charge on an electron.
- The spectrum from an X-ray tube has peaks which are characteristic of the metal from which the target anode is made. This line spectrum results from inner orbital electrons being ejected from the atoms of the target anode by incoming electrons. Other electrons replacing the ejected electrons lose a high-energy photon as they fall into the low energy inner orbitals.
- The penetrating properties of X-rays make them useful for imaging in medicine and industry because different densities of material affect the attenuation of the X-rays.
- The very small wavelength of X-rays makes them useful for diffraction experiments at the molecular or crystalline level.
- The photon energy of X-rays is related to their wavelength and their frequency according to $E_{ph} = hf = hc/\lambda$, where h is Planck's constant, f is the frequency of the radiation, c is the speed of light in a vacuum and λ is the wavelength of the radiation.
- The maximum photon energy, maximum frequency and minimum wavelength for an X-ray spectrum are given by $E_{ph(max)} = hf_{max} = hc/\lambda_{min} = q_e V$ where $E_{ph(max)}$ is the photon energy in joules, h is Planck's constant, f_{max} is the maximum frequency of the radiation in hertz, c is the speed of light in a vacuum, λ_{min} is the shortest wavelength of the radiation, V is the accelerating voltage of the X-ray tube, and q_e is the charge on an electron in coulombs (C).
- Synchrotron light, or radiation, is electromagnetic radiation that is emitted when charged particles travel at speeds close to that of light through a curved path under the influence of a magnetic field.
- Synchrotron light extends from the infrared region of the spectrum through to hard X-rays, and is highly collimated, travels in short pulses, has very high intensity, covers a broad spectral range, and is tunable and highly polarised.
- A collimated beam of X-rays incident upon a layer of atoms will be scattered. Bragg's law states that the beams will interfere.
- We can employ Bragg's law to analyse X-ray diffraction patterns. The structure of the unit cell defining a crystal can be identified though this process. This is called X-ray crystallography.

5.6 Questions

1 A typical X-ray tube in a dental clinic operates at 70.0 kV.
  - **a** Calculate the maximum kinetic energy of the accelerated electrons just prior to striking the target anode.
  - **b** Calculate the maximum speed of the electrons just prior to striking the target anode.
  - **c** Calculate the maximum photon energy of the emitted X-rays.

2 Consider a 150 kV X-ray tube being used in a factory to inspect welds in piping.
  - **a** Calculate the maximum photon energy of the emitted X-rays.
  - **b** Calculate the maximum frequency of the emitted X-rays.
  - **c** Calculate the minimum wavelength of the emitted X-rays.

3 The graph shows the spectrum obtained when a molybdenum target is bombarded with electrons accelerated with a voltage of 35 kV.

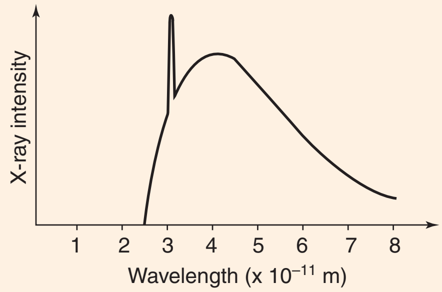


  - **a** What is the minimum wavelength of the emitted photons in metres?
  - **b** Why are there no X-rays with wavelengths less than your answer to part a?

4 What is the maximum frequency of the emitted X-ray photons in Question 3?

5 Explain the presence of the two peaks on the graph for Question 3.

6 The diagram shows an X-ray spectrum.



  - **a** What is the photon energy associated with the peak of the graph? Give you answer in joules and electronvolts.
  - **b** Explain why this peak is a particular characteristic of barium.
  - **c** Why are there no X-rays emitted with wavelengths of less than about 2.5×10^{-11} m?
  - **d** What is the maximum photon energy of the emitted X-rays?

7 Consider your answers to Question 6.
  - **a** Sketch a graph of X-ray intensity against photon energy for this data.
  - **b** Sketch a graph of X-ray intensity against frequency for this data.

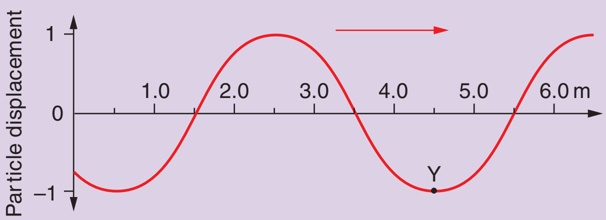
8 One type of target anode is a piece of tungsten embedded in a copper device. Why are these two particular metals used in such a device?

9 High-energy electrons travelling in a curved path under the influence of a magnetic field will emit photons with energies ranging from:
  - **A** microwaves to soft X-rays
  - **B** infrared to hard X-rays
  - **C** hard X-rays to gamma rays
  - **D** visible light to cosmic rays.

10 List three differences between X-rays generated in an X-ray tube and X-rays produced in a synchrotron.

Chapter 5 Review

The following information relates to questions 1–4.
The diagram shows the displacement of the air molecules in a sound wave from their mean positions as a function of distance from the source, at a particular time. The wave is travelling to the right at 346 m s^{-1}.



1 What is the wavelength of the sound wave?

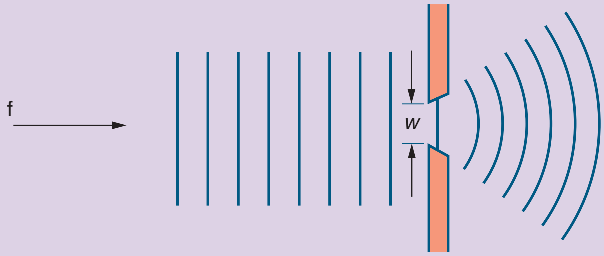
2 Which arrow describes the direction of motion of a molecule at point Y just after this particular time?

A → **B** ← **C** ↓ **D** ↑

3 Which arrow below describes the direction of transfer of acoustic energy by this wave?

A → **B** ← **C** ↓ **D** ↑

4 Which of the following properties of sound is independent of the source producing the sound?
  - **A** frequency
  - **B** amplitude
  - **C** speed



5 Sound waves of frequency f are being diffracted as they pass through a narrow slit of width w, as in the diagram above. The amount of diffraction can be increased (Choose one or more answers):

A by increasing f
B by increasing w
C by decreasing f
D by decreasing w.

The following information relates to questions 6 and 7.
A signal generator connected to a speaker produces sound waves that are directed into a tube closed at one end. The effective length of the tube is 85 cm.

6 What is the lowest frequency of sound that will produce resonance in the tube?

7 What frequency of sound will cause the tube to resonate at its third harmonic?

The following information relates to questions 8–11.

charge on electron: $e = 1.60 \times 10^{-19}$ C
mass of electron: $m_e = 9.11 \times 10^{-31}$ kg
mass of proton: $m_p = 1.67 \times 10^{-27}$ kg
Planck's constant: $h = 6.63 \times 10^{-34}$ J s
speed of light: $c = 3.00 \times 10^{8}$ m s^{-1}

8 A 500 W lamp directs a beam of yellow light, of wavelength 580 nm, on to a perfect reflecting surface of area 4.0 cm^2.
a Calculate the energy of each photon in the beam in:
i joules
ii electronvolts.
b What is the momentum of each photon in the beam?
c How many photons are incident on the reflecting surface each second?

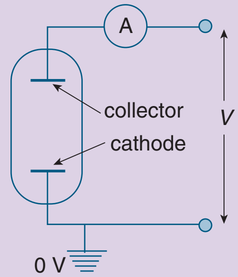
9 A photovoltaic cell consists of a metal surface coated with a thin layer of selenium. Light incident on the surface releases electrons, producing a small electric current. No electrons are ejected from the selenium surface. A 230 W lamp emits blue light, of wavelength 432 nm, so that the light beam is incident normally on a photovoltaic surface of area 5.0 cm^2. All the photons emitted by the lamp strike the photovoltaic surface.
a Calculate the energy in joules of a photon of blue light.
b What is the momentum of each photon?
c How many photons strike the photovoltaic surface each second?
d Assuming that only 0.001% of the incident photon energy is converted into electrical energy in the cell, determine the electrical power generated in the cell by the blue light.

10 When light is incident on a photosensitive metallic surface, electrons may be ejected from the surface. Which of the following is true of the speed of the ejected photoelectrons?
A It varies with the colour of the incident light.
B It varies with the intensity of the incident light.
C It varies with the speed of the incident light.
D It varies with the incident angle of the light.

11 Discuss how each of the following observed results of the photoelectric effect contradicts the wave model theory of light.
a Only certain frequencies of light can produce photoelectric emission from a photosensitive surface.
b A low-intensity light beam will produce photoelectric emission in the same time as a high-intensity light beam.
c A high-intensity light beam produces photoelectrons with the same maximum kinetic energy as a low-intensity light beam of the same frequency.

The following information applies to questions 12 and 13.
The cathode of the photoelectric cell shown below is coated with rubidium. Incident light of varying frequencies is directed onto the cathode of the cell and the maximum kinetic energy of the emitted photoelectrons is recorded. The results are summarised in the following table.

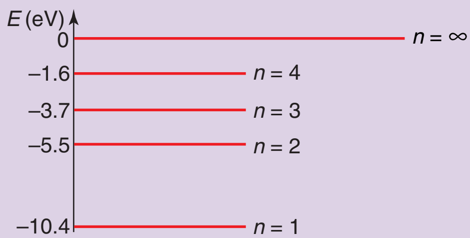
Frequency (Hz)	$E_{k(max)}$ (eV)
5.20×10^{14}	0.080
5.40×10^{14}	0.163
5.60×10^{14}	0.246
5.80×10^{14}	0.328
6.00×10^{14}	0.411
6.20×10^{14}	0.494



12 Plot a graph of the maximum kinetic energy of the ejected photoelectrons (J) as a function of the frequency of the incident radiation (Hz).
a Calculate the gradient of the graph.
b What does the gradient of the graph represent?
c Use the graph to determine the threshold frequency for rubidium.
d Will red light of wavelength 680 nm emit photoelectrons from a rubidium surface? Justify your answer.

13 Green light with a frequency of 5.60×10^{14} Hz is directed onto the cathode of the photoelectric cell.
a Using the threshold frequency determined above, together with the accepted value of Planck's constant, calculate the work function of rubidium in eV. Use this value of work function where required in all subsequent calculations.
b What is the kinetic energy (J) of the fastest photoelectrons emitted from the cathode?
c What is the momentum of the fastest photoelectrons emitted from the cathode?
d What is the magnitude of the minimum retarding potential difference that will prevent photoelectrons reaching the collector of the photoelectric cell?

The following information applies to questions 14 and 15.
The diagram shows the energy levels for atomic mercury.



14 An electron beam with energy of 8.80 eV passes through some mercury vapour in the ground state. Which one of the following photon energies could not be produced from this interaction?
A 1.8 eV **B** 4.9 eV **C** 3.7 eV **D** 6.7 eV **E** 3.9 eV

15 List all the possible photon frequencies that could result when a mercury atom returns to the ground state from the second excited state.

16 What is the shortest wavelength of light that can be absorbed by a hydrogen atom when an electron makes a transition from the ground state?

17 When a 10.2 eV photon is absorbed, a hydrogen atom will stay excited for approximately 10^{-8} s before the electron returns to ground state, emitting a photon of the same energy. Why then is this photon missing from the emission spectrum for hydrogen?

18 A 60.0 W lamp emits radiation of wavelength 3.00×10^{-6} m.
a How many photons per second is the lamp emitting?
b What type of light is the lamp emitting?

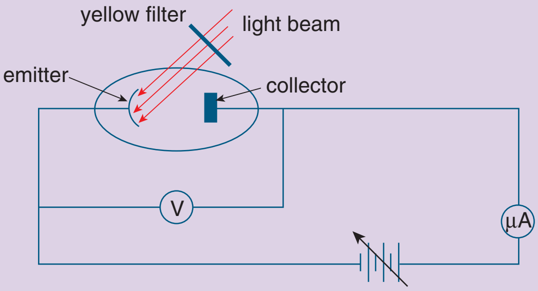
19 A beam of yellow light, 5.20×10^{14} Hz, is incident on a target which absorbs all the radiation falling on it. The beam delivers 1000 J of energy each second.

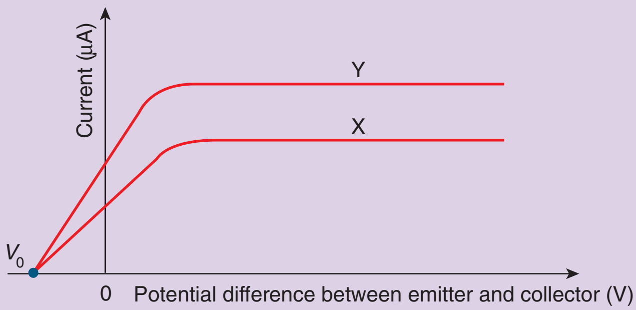
a How many photons are striking the target each second?

b What is the power of the beam?

The following information applies to questions 20–23.

Light passing through a yellow filter is incident on the cathode of the photoelectric cell in diagram (a). The reverse current in the circuit can be altered using a variable voltage. At the 'cut-off' voltage, V_{out}, the photoelectric current is zero. The current in the circuit is plotted as a function of the applied voltage in diagram (b).





20 Which of the following changes would result in an increase in the size of V_{out}?

A Replacing the yellow filter with a red filter

B Replacing the yellow filter with a blue filter

C Increasing the intensity of the yellow light

21 Which one of the following alternatives best describes the reason why there is zero current in the circuit when the applied voltage equals the cut-off voltage?

A The threshold frequency of the emitter increases to a value higher than the frequency of yellow light.

B The work function of the emitter is increased to a value higher than the energy of a photon of yellow light.

C The emitted photoelectrons do not have enough kinetic energy to reach the collector.

22 Which of the following descriptions of the graphs X and Y are correct?

A Both graphs are produced by yellow light of different intensities.

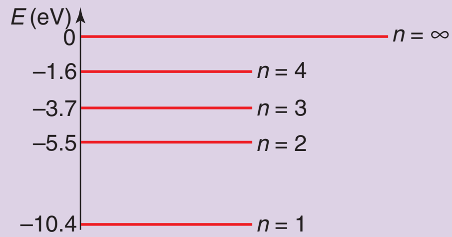
B Graph X is produced by yellow light while graph Y is produced by blue light.

C Each graph is produced by light of a different colour and different intensity.

23 The emitter of the photocell is coated with nickel. The filter is removed and a 200 nm light is directed onto the cathode. The minimum value of V_{out} that will result in zero current in the circuit is 1.21 V. What is the work function of nickel?

24 Describe three experimental results associated with the photoelectric effect that cannot be explained by a wave model theory of light.

25 The energy levels for atomic mercury are as follows.



Determine the frequency and wavelength of the light emitted when the atom makes the following transitions:

a $n = 4$ to $n = 1$

b $n = 2$ to $n = 1$

c $n = 4$ to $n = 3$

26 An electron is accelerated across a potential difference of 65.0 V.

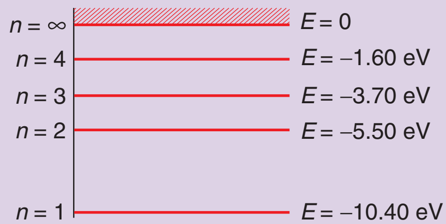
a What kinetic energy will it gain?

b What speed will it reach?

c What is the de Broglie wavelength of the electron?

The following information applies to questions 27–30.

The energy levels for atomic mercury are shown below.



27 a Calculate the energy of the photon that would be emitted if a mercury atom made a transition from the $n = 4$ state to the ground state.

b Calculate the wavelength of the photon that would be emitted if a mercury atom made a transition from the $n = 4$ state to the ground state.

c What is the minimum energy of an incident particle that could produce ionisation of a mercury atom in the ground state?

28 An electron beam of energy 7.00 eV passes through some mercury vapour in the ground state.

a Determine the energy of each electron in joules.

b What is the highest energy state that the mercury atom could be excited to?

c List all the possible electron energies that would be present in the emission spectrum.

d What is the shortest wavelength of light present in the emission spectrum?

29 A photon collides with a mercury atom in the ground state. As a result, a 30.4 eV electron is ejected from the atom. What was the wavelength of the incident photon?

30 Electrons of energy 4.00 eV travel through a glass tube containing mercury vapour.

a Will any photons be emitted from the mercury atoms in the tube? Justify your answer.

b Explain why there is a large increase in current through the circuit when electrons of energy 14 eV pass through the vapour.

c What wavelength light will be emitted from the tube when $V = 6.20$ V?

6

Matter, relativity and astronomy

Most people recognise Albert Einstein as the embodiment of genius. He was also a person of great humanity, one who cared deeply about the world in which he lived. At the end of World War II, for example, he was outspoken in his support for a ban on all nuclear weapons, as well as in denouncing all forms of intolerance and racism.

In 1905 Einstein published five papers in the prestigious German physics journal *Annalen der Physik*. Three of these papers changed forever the way in which we would understand the nature of our world. Two of these concerned the special theory of relativity and the other, the photoelectric effect. Relativity totally overturned our conceptions of the nature of space and time, linking them together into four-dimensional 'spacetime'. The photoelectric effect was a critical step towards the development of quantum theory, in which not only did light and matter become related in totally unexpected ways, but matter itself became curiously nebulous and somehow linked to the mind of the observer.

This photograph (opposite) shows the trails left by high-energy subatomic particles in a bubble chamber. The chamber contains a superheated liquid, usually helium or hydrogen, within a strong magnetic field. Charged particles fired into the chamber leave trails of gas bubbles in the liquid. Interactions between particles, or between particles and the atoms of superheated liquid, produce other particles, which can be identified by the curved paths they take after the interaction. Bubble chambers enable us to investigate electron–positron annihilation, among other things.

About ten billion years ago, there was a huge explosion in our corner of the galaxy—a fireball of truly astronomical proportions. A giant fast-burning star literally blew itself to pieces as a 'supernova'. In the short time before the cosmic dust was scattered, the energy released created temperatures so high that some of the star's hydrogen and helium atoms were fused together to form larger atoms—atoms such as carbon, oxygen, iron and uranium. Much later, some of that dust, interspersed with primordial hydrogen and pulled by the slow but relentless force of its own gravity, started to coalesce into a great whirling cloud. Most of the material eventually collapsed into the centre with such force that the temperature created ignited a new nuclear fire. Our Sun was formed. Some of the heavier material in the cloud settled into smaller clumps spinning around the Sun. The planets, including Earth, were born.

By the end of this chapter

you will have covered material from the study of matter, relativity and astronomy including:

- the problems with classical physics that led to Einstein's theory
- Einstein's postulates and their implications for relative motion
- the use of thought experiments to show that measurements of time and space depend on one's frame of reference
- the concept of time dilation and length contraction
- the development of our understanding of matter including bosons, leptons, mesons and baryons
- how bosons and the quark came to be accepted as the fundamental particles in nature
- the development of models of the solar system and the Universe
- the Copernican revolution and Galileo's contribution
- the properties of stars, distance, apparent and intrinsic brightness, and spectral type
- galaxies: types, distances, Hubble and red-shifts
- theories of the formation of the Universe, galaxies, stars and planets.

6.1 Extending our model of matter

Physicists get a new toy

In late 2008 the world held its breath as the Large Hadron Collider (LHC) was 'switched on' for the first time. The LHC is a particle accelerator built in a circular tunnel 100 metres underground, which accelerates both protons and lead ions in opposite directions with the intention of smashing them together. The resulting fragments and conditions are studied in order to discover new particles and to create the quark-gluon plasma that was thought to exist soon after the 'Big Bang'. It is thought that the LHC will help physicists to extend the standard model of matter and to discover new understandings of the fundamental forces in the Universe.

Physics file

What do young physicists who are working on the LHC do in their spare time? Well, some of them write a rap and publish it on YouTube. Check it out yourself by searching for 'large hadron collider rap'.

Figure 6.1
The Australian ANTARES tandem accelerator has carried out hundreds of carbon-dating procedures.

As far as you need to know . . .

In the past you may have heard your science teachers mention that the fundamental particles in nature are the proton, neutron and electron. What they should have said is that 'as far as you need to know' the fundamental particles in nature are the proton, neutron and electron. The actual situation is far more quirky, strange and charming than this. As a spin-off from the investigations into the nucleus that were conducted during the latter stages of World War II, particle physicists began to recognise that as the energy of the bombarding particle was increased, new particles were being formed. After the war, physicists were building particle detectors and then waiting for high-energy cosmic rays from space to smash into their targets, in order to see what nuclear fragments they could identify. It was recognised that in order to probe more effectively into the nucleus, physicists needed to build more powerful machines to accelerate protons and electrons to energies high enough to form the particles that they were predicting would exist and to expose the fundamental forces in nature.

Physics file

As far as we know, the four fundamental forces in nature, in order from highest strength to lowest, are the strong nuclear force, the electromagnetic force, the weak nuclear force and the gravitational force. The relative strengths of these forces, when compared to the strong nuclear force, are $1 : 10^{-2} : 10^{-6} : 10^{-38}$.

Resolving small particles

The ability to resolve, or image, small particles is related to the wavelength of the light that is used to illuminate the particles. From our understanding of diffraction, we know that when waves have a wavelength that is greater than the width of the particle, they will be diffracted around the particle. Significant scattering occurs for wavelengths of approximately the same size or smaller than the object; it is the scattered waves that provide us with the image data. Visible light is limited to the size of particle it can resolve as its wavelength is about 600 nm; higher frequency, and therefore smaller wavelength, X-rays can be used to resolve atoms within a crystal structure to 0.1 nm. If we change from electromagnetic waves to matter, we can start to use even smaller wavelengths to probe the nucleus. From de Broglie's work on matter waves, we know that high-energy electron beams have very short wavelengths, which can then be used to resolve particles about the size of a nucleus. The de Broglie wavelength is calculated by using:

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

where λ is the wavelength (m), h is Planck's constant (J s), c is the speed of light (m s^{-1}) and E_k is the kinetic energy of the particle (J).

✓ Worked Example 6.1A

The Australian National Tandem Accelerator for Applied Research (ANTARES) is a linear accelerator that can produce electrons with 10.0 MeV of kinetic energy. What is the minimum size of particle that ANTARES can resolve?

Solution

$c = 3.00 \times 10^{8}\ \text{m s}^{-1}$

$h = 6.63 \times 10^{-34}\ \text{J s}$

$E_k = 10.0 \times 10^{6}\ \text{eV}$

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

$$= \frac{(6.63 \times 10^{-34})(3.00 \times 10^{8})}{(10.0 \times 10^{6})(1.60 \times 10^{-19})}$$

$$= 1.24 \times 10^{-13}\ \text{m}$$

Note that atoms are approximately 10^{-10} m in diameter, while nuclei are approximately 10^{-14} m, so ANTARES can resolve an atom, but not its nucleus.

Force fields or particle exchange

Before we start delving into the nucleus, we need to change the way in which we think about force. Previously we have been thinking and talking about force fields as regions of space in which a force is experienced by a particle, with specific properties that are related to the field. For example, an electric field will cause a positively charged particle to move under the influence of an electromagnetic force. We are also familiar with the idea that an electric field is actually a component of electromagnetic radiation, which propagates as a wave and is able to spread the electric force out into space. But if we consider the nature of electromagnetic radiation, the notion of wave–particle duality says that light also travels as a particle called a photon. Therefore, the opposing electric force that a pair of charged particles experiences could result from either an electromagnetic field created by each charged particle or by the exchange of photons between the two charged particles.

To continue with our exploration of the nucleus in greater detail, it is necessary to consider that a force between two particles results from an exchange of a force-carrying particle between the two. If two charged particles are approaching each other, then according to this model photons are being continuously exchanged between the two, causing a repulsive force on each charged particle. As the particles get closer, a greater number of exchanges take place and a stronger repulsive force is applied.

Feynman diagrams

The diagram in Figure 6.2 illustrates the concept in a simplistic way, but it would be more helpful to be able to visualise the interactions between particles and their force particles if we use diagrams invented by the American physicist Richard Feynman.

Figure 6.2
Two electrons approaching each other will exchange photons and will therefore apply a repulsive force on each other.

Feynman based his diagrams on a theory called quantum electrodynamics (QED), and used them to help explain the interactions between charged particles. In his diagrams, time progresses upwards on the *y*-axis while position is along the *x*-axis. These axes are dimensionless as they are meant to indicate the types of interactions that occur between particles and not relative distances or times for the interactions to occur. Arrows indicate the type of particles in the interaction, with normal matter arrows pointing up the page and anti-matter arrows pointing down the page. Note that these arrows do not indicate the direction in which the particles travel; they are simply meant to differentiate between the two forms of matter. Time always progresses upwards.

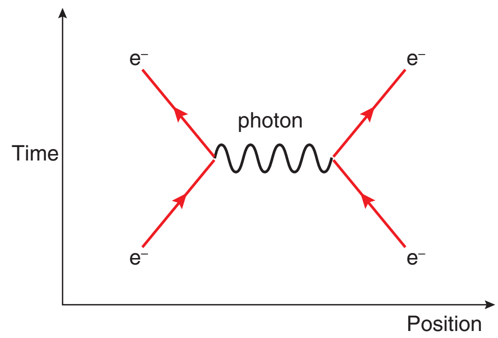


Figure 6.3

The Feynman diagram for the interaction between two electrons.

From our previous example in Figure 6.2, in which two electrons exert a repulsive force on each other, the Feynman diagram is shown in Figure 6.3. The Feynman diagram shows a photon being exchanged between the two electrons; notice that the two electrons change their direction in a way that suggests that they are being repelled. The photon exchanges energy and momentum from one electron to the other, It is important to note that this photon cannot be detected, as a normal photon can, as it exists for a very short period of time. These photons are known as *virtual photons*. The photon in this interaction is the force carrier, or is said to mediate the electromagnetic force. The Japanese physicist Hideki Yukawa postulated that as the photon mediates the electromagnetic force, then there must be a particle that mediates the strong nuclear force that acts between nucleons within the nucleus. Yukawa's force carrier, named the meson, was discovered in 1947, but was later found to be just one of a number of particles that mediate the strong force. These particles were now grouped under a larger category called *gluons*, because they glue the nucleus together. This discovery marked the beginning of the model of matter that we now call the standard model.

Physics in action — Breaking the law!

For a proton at rest to create a strong nuclear force-carrying particle like a meson, it would have to create the particle from nothing, but we know that this would violate the law of conservation of energy. A proton at rest has no kinetic energy, and therefore has no energy to make the new particle. However, Heisenberg's uncertainty principle allows for brief non-conservation of energy if the time period over which it is not conserved is very short. As the meson travels at very close to the speed of light (3.00×10^8 m s^{-1}) and the maximum distance it travels between two nucleons is very small (1.50×10^{-15}), then the time that the meson exists falls well within the time limit for uncertainty.

$d = 1.5 \times 10^{-15}$ m $\qquad v = \dfrac{d}{\Delta t}$

$v \approx 3.00 \times 10^{8}$ m s^{-1} $\qquad \Delta t = \dfrac{1.50 \times 10^{-15}}{3.00 \times 10^{8}} = 5.00 \times 10^{-15}$ s

The four forces in nature and their mediating particles

The four known fundamental forces or interactions in nature are the strong nuclear force, the electromagnetic force, the weak nuclear force and gravity. Two of the forces have particles that mediate their force field, but what of the weak nuclear and gravitational forces?

The electromagnetic force has its force-carrying particle, the photon, and the strong nuclear force has the gluon. The weak nuclear force also has its force-carrying particles, known as the W^+, W^- and the Z^0 particles, which were discovered in 1983 at CERN. The search for the wave-particle moderator of the gravitational force, called the graviton, is underway at places like the Australian International Gravitational Research Centre at University of Western Australia and their Gravitational Observatory located near the Gingin.

Matter and anti-matter

For each particle that exists, there is an anti-matter particle to match; for example an electron has its anti-matter particle, the positron. The positron has many of the same properties including mass, but it has a positive charge as opposed to the negative charge of the electron. Other features of the proton are also opposite to the electron's properties. Once the positron was discovered, other anti-matter particles began to be uncovered such as the anti-proton and the anti-neutron. Anti-matter is extremely rare on Earth and is greatly outnumbered in the Universe by regular matter, which is a puzzle. When normal matter interacts with its anti-matter equivalent, the two particles will convert all of their mass into energy ($E = mc^2$) in a process called *annihilation*.

✓ Worked Example 6.1B

Calculate the energy released when an electron meets a positron and they are annihilated.

Solution

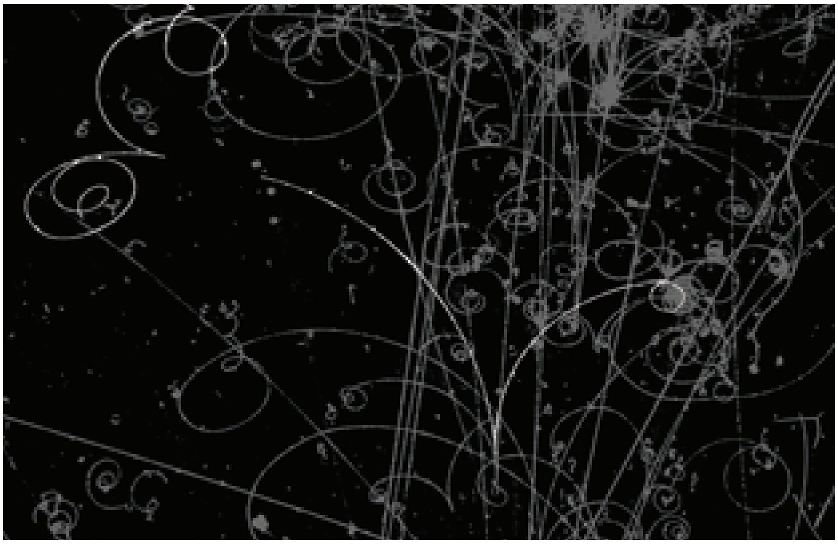
An electron has a mass of 9.11×10^{-31} kg; the anti-matter positron has many properties that are opposite to the electron, but it has the same mass, therefore the energy created when the two meet is:

$m_{e^-} = 9.11 \times 10^{-31}$ kg

$m_{e^+} = 9.11 \times 10^{-31}$ kg

$$E = mc^2$$
$$= (9.11 \times 10^{-31} + 9.11 \times 10^{-31})(3.00 \times 10^8)^2$$
$$= 1.64 \times 10^{-13} \text{ J}$$

This energy is usually in the form of an emitted photon with a wavelength of 1.21×10^{-12} m, which places it in the gamma ray section of the electromagnetic spectrum. This is also the minimum wavelength photon that could create an electron–positron pair; a higher energy photon would result in the pair having some kinetic energy.



Physics file

Located near Gingin, 80 km north of Perth, is the Australian International Gravitational Observatory (AIGO). The facility is linked to other similar observatories around the world to increase its resolution and sensitivity to incoming gravity waves. Two 80-metre tubes capable of maintaining a high vacuum are aligned at 90° to each other, with an extremely flat mirror at each end. A laser beam is sent down one tube and back again off the mirror; it then reflects down the other tube and back again. When the two reflected laser beams interfere with each other, an interference pattern is created. If a gravity wave passes through the observatory, it is thought that space itself will distort in time with the gravity wave. If this occurs, then physicists should expect to see that the interference pattern changes as one arm of the observatory will be slightly longer than the other arm.

Figure 6.4
The Australian International Gravitational Observatory near Gingin, north of Perth. The nearby Gravity Discovery Centre is worth visiting.

Practical activities

34 Detecting radiation with a Geiger–Muller tube
35 The diffusion cloud chamber

Figure 6.5
A bubble chamber photograph of an electron-positron pair production and annihilation. A gamma ray enters from the bottom and produces an electron which curves to the right, and a positron which curves to the left. The positron track stops abruptly as it strikes an electron in the vapour and annihilates. One of the photons produced in the annihilation then goes on to produce another electron-positron pair, but with lower kinetic energy. The oppositely charged particles are forced to curve in different directions by a magnetic field directed downwards.

Figure 6.6

Satyendra Bose, the Indian physicist whose work on quantum statistics was extended by Einstein to create the Bose–Einstein statistics system that describes particles with an integral spin. These particles are called bosons in honour of his work.

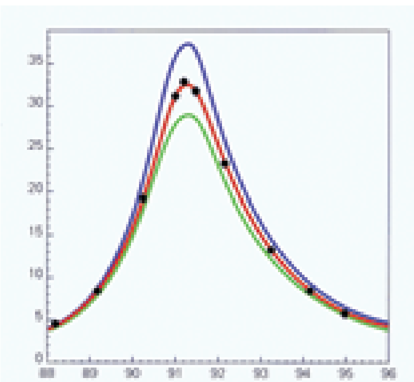


Figure 6.7

These three lines model the existence of two (green), three (red) or four (blue) neutrinos. The data from millions of Z particles, observed by the ALEPH (Apparatus for LEP Physics) experiment at CERN, produced the black dots. There is a strong correlation between the data and three neutrino model.

Figure 6.8

Neils Bohr, Hideki Yukawa, his wife Sumiko Yukawa and Robert Oppenheimer. Yukawa was awarded the Nobel Prize in Physics in 1949 for his prediction of the existence of mesons.

Classifying these new particles

After Yukawa discovered the new pi meson, researchers began to uncover hundreds of new particles that each had unique properties. By grouping particles with similar properties, a system of classification evolved that was based on the type of interactions that correspond to the particles.

Bosons

This group of particles mediate the strong nuclear, electromagnetic and weak forces. They include the gluons, photons and the W and Z particles. Recall that we are considering these force fields in their particle nature, according to the notion of wave–particle duality.

Leptons

This group of particles interact by exchanging W and Z bosons, which mediate the weak nuclear force. Leptons that carry a charge can also interact by exchanging photons, which mediate the electromagnetic force, but they do not interact via the strong force carriers. These particles include the electron, muon and tau particles, as well as their corresponding neutrinos and the anti-matter opposites of these six particles. Recall from Unit 2A that an electron ejected from the nucleus in a beta minus decay will always be emitted along with an anti-neutrino, and an anti-matter positron is emitted with a normal neutrino.

Hadrons

This group of particles is further categorised into two groups is, the baryons and the mesons, based on a property called baryon number (B), described later in this section. The common feature of the hadrons is that they all interact by exchanging gluons, which are the particles that mediate the strong nuclear force. Hadrons that carry a charge can also interact by exchanging photons, but the effect of the gluons far outweighs any other force mediating particle.

Mesons

These are particles that have been assigned a baryon number of zero. In this group are many particles and their anti-particles, for example the pion (π^+), anti-pion (π^-) and pi-zero (π^0), the kaon (K^+) and anti-kaon (K^-), and the eta (η^0). Note that the anti-particle of the eta is considered to be itself.

Baryons

These are particles that have been assigned a baryon number of 1 for normal matter or –1 for anti-matter. In this group are the familiar proton (p^+) and anti-proton (p^-), neutron (n) and anti-neutron ($\overline{n}$), along with hundreds of other particles and their anti-particles, for example the lambda-zero (Λ^0) and anti-lambda-zero ($\overline{\Lambda}^0$), sigma-plus (Σ^+), sigma-zero (Σ^0) and sigma-minus (Σ^-), the xi-zero (Ξ^0) and omega-minus (Ω^-).

As far as they know!

As far as physicists know, all six leptons are considered to be fundamental particles as there appears to be no internal structure to them and their size appears to be unmeasurable. Hadrons on the other hand appear to have an internal structure, indicating that they are actually made of smaller particles. These smaller particles must then be considered as the fundamental particles in nature.

The physicists Murray Gell-Mann and George Zweig first suggested that every hadron was made up of combinations of the fundamental particles that they called quarks. Their initial theory had three quarks, but there are now considered to be six quarks, just as there are six leptons. This theory has gained general acceptance by the physics community as the standard model of matter.

What flavour quark would you like?

High-energy particle physicists seem to be a creative bunch; not only do they choose unusual names for their new particles but they also choose unusual ways to describe them and their properties. The six different types of quarks are referred to as the six flavours of quarks.

The six flavours of quarks are the up (u), down (d), strange (s), charmed (c), bottom (b) and top (t). The last four quark names also apply to new quantum numbers and their conservation laws called strangeness (S), charm (c), bottomness (b) and topness (t). Quarks have their anti-matter opposites called anti-quarks that have the opposite sign for all of their quantum numbers. Quarks have properties of baryon number, spin and charge, which must add together to give the total baryon number, spin and charge of the hadrons they combine to make. Quarks also have another property called colour charge, which is similar to electric charge, and distinguishes between two quarks of the same flavour, but with different spin. The three colours of quarks are red, green and blue and the colours for the anti-quarks are anti-red (cyan), anti-green (magneta) and anti-blue (yellow). Quarks must combine so that their colour charge property adds to equal white (for example red + green + blue = white) but red + cyan also equals white as do green + magenta and blue + yellow!

Physics file

The name 'quark' was taken by Murray Gell-Mann from the book *Finnegans Wake* by James Joyce, which contains the line 'Three quarks for Muster Mark'.

Figure 6.9
This is one of the first pieces of photographic evidence of the presence of a neutron. It shows the cloud-chamber track of a proton that has been struck by a neutron, the path of the neutron is not seen as it has no charge. This photo was taken by French physicists Irene Curie and Frederic Joliot in 1932.

table 6.1 The various properties of quarks

Name	Symbol	Charge (Q)	Baryon number (B)	Strangeness (S)	Charm (c)	Bottomness (b)	Topness (t)
Up	u	$+\frac{2}{3}$ e	$\frac{1}{3}$	0	0	0	0
Down	d	$-\frac{1}{3}$ e	$\frac{1}{3}$	0	0	0	0
Strange	s	$-\frac{1}{3}$ e	$\frac{1}{3}$	−1	0	0	0
Charmed	c	$+\frac{2}{3}$ e	$\frac{1}{3}$	0	+1	0	0
Bottom	b	$-\frac{1}{3}$ e	$\frac{1}{3}$	0	0	−1	0
Top	t	$+\frac{2}{3}$ e	$\frac{1}{3}$	0	0	0	+1

How to make a baryon

Baryons, which include the proton and neutron, consist of three quarks; for example, the proton is made up of two up quarks and a down quark (uud). Each of these quarks must have a different colour charge of red, green and blue. But they must also combine to have a total charge of the

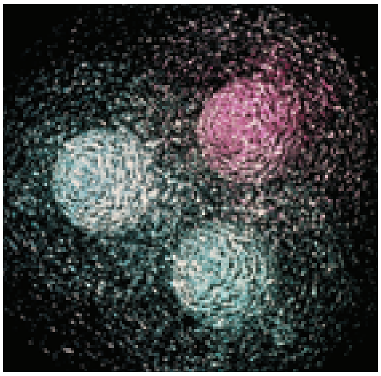


Figure 6.10
An artist's representation of the three quarks that make up a proton (uud). Notice that the three quarks have a colour representing the colour charge responsible for the strong nuclear force. The quarks are surrounded by a quantum cloud of virtual particles.

proton, which is one 'fundamental unit' ($+1.60 \times 10^{-19}$ C). Therefore, quarks must have some charge less that the fundamental unit. In fact, quarks have a charge of either $+\frac{2}{3}$ e (up, charmed and top) or $-\frac{1}{3}$ e (down, strange and bottom), where e is the former fundamental unit of charge. The proton charge is therefore made up of $\left(+\frac{2}{3}\right)+\left(+\frac{2}{3}\right)+\left(-\frac{1}{3}\right)=\left(+\frac{3}{3}\right)=+1$. A neutron is made up of two down quarks and an up quark (ddu) of different colours, which equates to a charge of $\left(-\frac{1}{3}\right)+\left(-\frac{1}{3}\right)+\left(+\frac{2}{3}\right)=\left(\frac{0}{3}\right)=0$. Note that the baryon number of each quark is $+\frac{1}{3}$ so three quarks add to give a baryon number of +1; anti-quarks have a baryon number of $-\frac{1}{3}$.

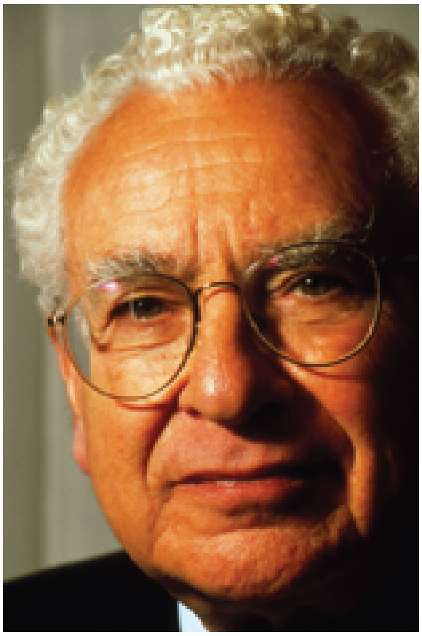


Figure 6.11
Murray Gell-Mann entered university at the age of 15, had his doctorate at 22 and introduced the concept of strangeness at the age of 24. At 35, with George Zweig, he proposed that nucleons were themselves made up of quarks and at the age of 40 he won the 1969 Nobel Prize in Physics.

table 6.2 Table of baryons and their quarks

Name	Symbol	*B*	*S*	*c*	*b*	*t*	Quarks
Proton	p	+1	0	0	0	0	uud
Anti-proton	$\bar{p}$	−1	0	0	0	0	$\overline{uud}$
Neutron	n	+1	0	0	0	0	udd
Anti-neutron	$\bar{n}$	−1	0	0	0	0	$\overline{udd}$
Lambda-plus	Λ^+	+1	0	+1	0	0	udc
Lambda-zero	Λ^0	+1	−1	0	0	0	uds
Sigma-plus	Σ^+	+1	−1	0	0	0	uus
Sigma-zero	Σ^0	+1	−1	0	0	0	uds
Sigma-minus	Σ^-	+1	−1	0	0	0	dds
Xi-zero	Ξ^0	+1	−2	0	0	0	uss
Xi-plus	Ξ^+	+1	−2	0	0	0	dss
Omega-minus	Ω^-	+1	−3	0	0	0	sss

Interactive tutorials
Atomic stability
Radioactive decay and half-life

How to make a meson

Mesons consist of a quark and an anti-quark pair, for example the pion-plus (π^+) consists of an up quark and an anti-down quark ($u\bar{d}$). Therefore, the charge of the pion-plus is $\left(+\frac{2}{3}\right)+\left(+\frac{1}{3}\right)=\left(+\frac{3}{3}\right)=+1$.

table 6.3 Table of mesons and their quarks

Name	Symbol	*B*	*S*	*c*	*b*	*t*	Quarks
Pion-plus	π^+	0	0	0	0	0	$u\bar{d}$
Pion-minus	π^-	0	0	0	0	0	$\bar{u}d$
Kaon-plus	K^+	0	+1	0	0	0	$u\bar{s}$
Kaon-minus	K^-	0	−1	0	0	0	$\bar{u}s$
Rho-plus	ρ^+	0	+1	0	0	0	$u\bar{d}$
Rho-minus	ρ^-	0	−1	0	0	0	$\bar{u}d$
phi	φ	0	0	0	0	0	$s\bar{s}$
D-plus	D^+	0	0	+1	0	0	$c\bar{d}$
D-zero	D^0	0	0	+1	0	0	$c\bar{u}$
D-plus-s	$D^+{}_s$	0	+1	+1	0	0	$c\bar{s}$
B-minus	B^-	0	0	0	−1	0	$b\bar{u}$
Upsilon	Υ	0	0	0	0	0	$b\bar{b}$

Strong nuclear force and the colour charge

According to the theory of quantum chromodynamics (QCD), the force that binds the quarks together in a hadron is called the colour force and it acts between the colour charge of the quarks. The strong nuclear force between nucleons or between any hadrons is therefore a result of the colour force between the quarks within each of the hadrons. So the attraction between one proton and another, within a nucleus, is a result of the attraction between the quarks contained within each proton. This leads to the gluon particles that are exchanged in the strong nuclear force being the same particles that exchange the colour force.

Physics file

Our chances of ever detecting a single quark are very unlikely; they are so tightly bound that an enormous amount of energy would be required, so that more particles would be created out of the kinetic energy of the bombarding particles. So instead of splitting a meson, we would just create more mesons and baryons.

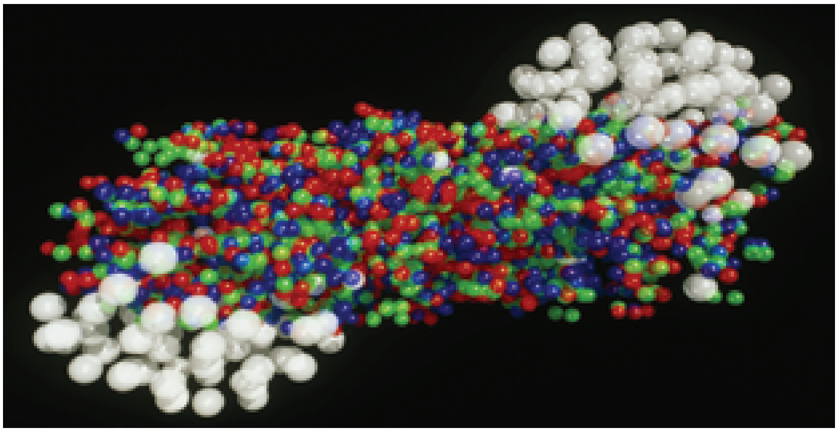


Figure 6.12
Scientists at the Large Hadron Collider at CERN hope to produce real interactions similar to this computer-generated model of two lead nuclei colliding. The white spheres are complete nucleons, while the red, green and blue spheres make up the quark-gluon plasma that has formed due to the energy of the collision. It is thought that this quark-gluon plasma was the state of matter in the Universe just after the Big Bang.

6.1 SUMMARY Extending our model of matter

- With higher frequency waves we are able to resolve images in greater detail. Matter waves have a higher frequency than ordinary light and cosmic rays; the higher the energy of the matter, the higher the frequency of the matter wave.
- An electric force can be considered to result from an electric force field (wave nature) or a photon particle exchange (particle nature).
- Feynman diagrams can illustrate the interactions between particles and the force carriers that cause the interactions.
- The four fundamental forces and the particles that mediate the force are the strong nuclear force (gluons), the electromagnetic force (photons), the weak nuclear force (W^+, W^- and Z^0), and the gravitational force (graviton).
- For every matter particle, there is an anti-matter particle; if the two ever meet, annihilation occurs.
- Matter is now classified as bosons (force-carrying particles), leptons (do not interact via the strong force carriers) and hadrons, which can be further divided into mesons (interact via the weak force carriers) and baryons (interact via the strong force carriers).
- Protons and neutrons are no longer considered to be fundamental particles. The new fundamental particles are the bosons (force carriers), leptons (includes the electron) and quarks (which make up protons and neutrons). Hundreds of other particles have been discovered that are made up of combinations of these fundamental particles.
- Quarks have a variety of properties including charge, baryon number, colour charge, strangeness, charm, bottomness and topness.
- A quark's charge is less than the fundamental charge unit: up, charmed and top = $+\frac{2}{3}$e, and down, strange and bottom = $-\frac{1}{3}$e.
- Mesons consist of a quark and an anti-quark, one of a particular colour charge and the other of its complimentary colour charge, while a baryon is made of three quarks, one of each colour charge.
- The force between colour charges of quarks is responsible for keeping quarks together in a meson or a baryon. It is this colour force that causes the strong nuclear force between mesons and baryons.

6.1 Questions

1 What is the relative strength of the fundamental forces of nature, when compared to the gravitational force?

2 Calculate the maximum size of particle that the following waves could resolve.
  a Visible light of wavelength 5.00×10^{-7} m
  b X-rays with a frequency of 3.00×10^{-11} m
  c A proton travelling at 0.650 times the speed of light

3 Draw the Feynman diagrams for the following interactions.
  a Electrostatic repulsion between a proton and proton (not within a nucleus)
  b Strong attraction between a proton and a neutron within a nucleus
  c A neutron changing into a proton by exchanging a pion with a proton that becomes a neutron

4 Calculate the energy produced in the complete annihilation of:
  a a proton and an anti-proton
  b a neutron and an anti-neutron.

5 Determine the kinetic energy of the particles produced if the following photons produce a matter anti-matter pair.
  a A 1.20 MeV photon produces an electron and a positron.
  b A photon of frequency 4.541×10^{23} Hz produces a proton and an anti-proton.
  c A photon of wavelength 6.059×10^{-16} m produces a neutron and an anti-neutron.

6 What combination of quarks exist in the following particles?
  a A neutron
  b An anti-proton
  c A lambda-zero (Λ^0)
  d A D^0 meson with $Q = 0$, $B = 0$, $S = 0$, $c = +1$

7 Which particle is represented by the following quark combinations?
  a uud
  b $d\bar{u}$
  c $s\bar{u}$
  d sdd

6.2 Einstein's special theory of relativity

At the end of the 19th century, physicists thought their theories described just about everything. Anything that moved—from atoms to planets—seemed to obey Newton's laws as precisely as could be measured. Maxwell's equations for electromagnetism were the equivalent for electrical phenomena, and chemistry was just a matter of developing a better understanding of the way in which the positive and negative particles that made up the atom obeyed these principles.

There were a few small problems, however. Maxwell's equations, which had accurately predicted that light was an electromagnetic wave, seemed to suggest that the speed of light would not obey the well-proven laws of relative motion first suggested by Galileo. Furthermore, the spectrum of light coming from 'excited' atoms appeared in sharp colour lines instead of the continuous rainbow band that theory predicted.

As it turned out, the first of these difficulties eventually led Einstein to put forward a radical new theory of space, time and relative motion, and the second led to the quantum theory put forward by Bohr and others. It is Einstein's special theory of relativity that we will investigate in this study.

Radical thinking

The story of physics is a story of radical thinkers. Put yourself in Copernicus's shoes for a moment. You have come up with what, to most people, seems a crazy idea: that the Earth actually moves around the Sun. How can you possibly convince people that this is the case? If you watch the Sun and the stars, how can you seriously suggest that it is not them but the Earth that is moving? And yet, on the basis of his careful study of motion in the heavens, Copernicus persisted with his idea—and we all know the result.

Sixty years later, Galileo not only tackled those who thought they knew the Earth was the centre of the Universe, but argued against other ideas of the great Greek philosopher Aristotle, whose views had been dogma for the past 2000 years. Aristotle had said that things fall to the Earth at a rate dependent on the amount of the 'earth element' they contain. So a large rock with more earth must fall faster than a leaf which contains fire, air and water as well as earth. Observations of falling rocks and leaves clearly support this idea. Aristotle also said that a large rock with more earth would fall faster than a small rock. Galileo was not convinced about this, so he actually did experiments and pondered on various types of motion. Eventually he put forward a radical new idea, one which is essential to our story of Einstein's relativity, itself one of the most radical ideas of all. We return to Galileo's idea shortly, but first let's catch a glimpse of Einstein's.

Moving through spacetime

We are very used to moving through time and space. As we sit in a chair reading this, we are moving through time at the rate of 24 hours every day, but we are not moving through space (unless we are in a train). If we take a fast trip—say, to the other side of the world—we have moved through space as well as time. Einstein's radical idea was that travel through space and travel through time are interrelated. In a sense, the faster we travel through space, the less we travel through time. It is as though we move through what he called *spacetime* at a constant rate, but the amount of space and the amount of time depend on who measures them.

As we watch a space traveller, it turns out that our measurements of her time elapsed and distance travelled do not agree with her measurements. We find that her time is going more slowly than ours. However, she sees that our clocks are going slow! So who is right? The answer is that we *both* are. If that sounds paradoxical, it is because we find it hard to accept the idea that time and space are *relative*. But that is what Einstein's theory of relativity is all about.

In what follows, we will discuss rocket ships travelling near the speed of light, and the malleable nature of the relationship between time and space. These are certainly some of the fascinating aspects of Einstein's incredible theory. However, relativity is not just about somewhat impractical high-speed space travel. What is often not realised is that Einstein started his famous 1905 paper with a discussion on the forces between objects moving at speeds of only millimetres per second. These are electrons moving in wires, the electric currents which create the magnetic forces that turn all the world's electric motors. Perhaps one of the most fascinating aspects of Einstein's relativity is that without it we cannot understand one of the most common and useful forces we find around us!

Galileo's principle of relativity

One of Galileo's most radical ideas was the principle of inertia, often referred to as Newton's first law. Imagine being so bold as to suggest that the natural state of objects is not to be at rest, but in a state of uniform

Figure 6.13
Galileo may or may not have dropped things from the Leaning Tower of Pisa, but his conclusions about falling objects changed history.

Physics file

A *frame of reference* is just physicists' way of describing a particular system of measurement coordinates. Our usual frame of reference is the Earth's surface. When we say a car was doing 115 km h^{-1} in a 100 km h^{-1} zone, we are assuming that the car's velocity was measured by a police officer who was at rest relative to the road, or maybe by a police officer who added the velocity of his police car to the speed obtained from his radar gun. In the frame of reference of the police car, which happened to be following at 100 km h^{-1} relative to the road, and in which the radar gun was mounted, the speeding car was travelling at 15 km h^{-1}.

In the larger scale of things, the Earth's surface is not a very satisfactory frame of reference. It is rotating, so different parts of it are travelling at different velocities relative to other parts. At any time, for example, Londoners are travelling at about 2000 km h^{-1} relative to people in Perth. (Think about it: If the Earth was transparent, how would we see London moving relative to the fixed stars beyond?) Of course we could equally say that Perth residents are travelling at 2000 km h^{-1} relative to Londoners, albeit in the opposite direction. The important thing to remember about frames of reference is that, while some might be particularly useful, none is any better, or more fundamental, than any other.

motion, a state of rest simply being a special case of uniform motion. This was quite contrary to the beliefs of the followers of Aristotle, who said that a force is necessary for motion. Without a force, they said, all motion would cease.

Hidden in the principle of inertia was an idea we now call the *Galilean principle of relativity*, which we can sum up by saying that there is nothing special about a velocity of zero. In Aristotle's world, motion only occurred when a force caused it and so where there was no force, there was no motion (i.e. zero velocity). Zero velocity was something very special. However, in the world of Galileo and Newton, zero velocity is no different, in principle, to any other velocity. A force (more particularly an impulse) that changes the speed of a ball from zero to 50 m s^{-1} will also change it from –25 m s^{-1} to +25 m s^{-1}, or indeed from 1000 m s^{-1} to 1050 m s^{-1}. That a force *changes* a velocity, rather than causes it, was a profound shift in thinking that was difficult to accept. In fact, even now, beginning physics students often find it difficult to grasp the implications of this idea. Aristotle's view was based on everyday experience, and despite the TV images we see of astronauts floating around in spacecraft, everyday experience still tends to tell us that without a force there is no motion. Indeed, despite 300 years of Newtonian physics, surveys show that when it comes to ideas about motion, the average person still thinks like Aristotle!

Interactive tutorial

Relative velocities

Figure 6.14

How fast is the astronaut moving? Relative to the space shuttle, he is moving very slowly, but relative to us on Earth he is moving at around 8000 m s^{-1}. In the frame of reference of the Sun he is moving at around 30 000 m s^{-1}.

If a force only causes a *change* of velocity, then it doesn't really matter from whose perspective we measure velocity. Velocity is always measured *relative* to some particular coordinate system or *frame of reference* (or sometimes just *frame* for short). Most measurements we make are relative to the Earth's surface, but this does not have to be the case. As long as we make our measurements from a frame of reference that is moving at constant velocity, any measurements of *changes* of velocity will agree with those made by observers in other steadily moving frames of reference. This is basically what Galileo's principle of relativity is about.

GALILEO'S PRINCIPLE OF RELATIVITY states that all motion is relative to some particular frame of reference, but there can be no frame of reference that has an absolute zero velocity.

To appreciate the principle of relativity, think about a simple situation in which you are moving along at a steady velocity. Imagine riding in a very smooth train speeding along a straight track. Inside the train, you can (provided it is not annoying other passengers) have a ball game with your friend at the other end of the carriage. As you throw the ball back and forth, you find that there is absolutely no difference between this game and one in the same carriage when it is stopped at a station. In fact, if you pull down the blinds, you would not even know you were moving. (This is a very quiet train!) In your frame of reference of the train, all the normal laws of physics about throwing and catching balls work just as they do on the school oval.

Another of your friends, Clare, is watching this ball game from beside the track. Clare sees you, and the ball, travelling at very different speeds to those you perceive. However, every time the ball is caught and thrown back, the *change* of speed she measures agrees exactly with the *change* that you measure. This is best illustrated by an example.

✓ Worked Example 6.2A

Angela and Bill are throwing a ball, of mass 200 g, back and forth in a train moving along at a steady speed of 30 m s^{-1} (108 km h^{-1}), as shown in Figure 6.15. Angela, who is at the front of the carriage, throws the ball at 10 m s^{-1}. Bill catches it and throws it back at the same speed.

a Clare, who is watching this game from alongside the tracks, has video equipment with which she can determine the various velocities involved. At what velocity does she find the ball was moving?

b Bill catches and throws the ball back in one steady movement that takes 2 s, during which he applies a constant force. What was this force from his point of view, and from Clare's point of view?

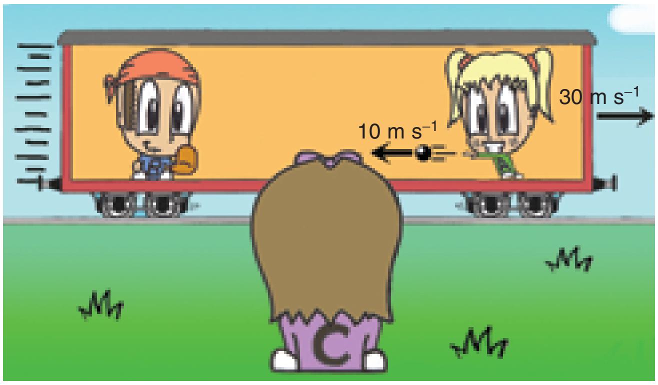


Figure 6.15
Angela has thrown the ball to Bill at 10 m s^{-1}, but how fast is the ball really moving?

Solution

We will choose the direction of the train's motion as positive and so Angela throws the ball with a velocity of −10 m s^{-1} in the frame of reference of the train.

a Clare sees Angela throw the ball backwards relative to the train's motion. Its velocity will be:

$\longleftrightarrow$ $- \quad +$

$v_{\text{ball–train}} = -10 \text{ m s}^{-1}$

$v_{\text{train–Earth}} = +30 \text{ m s}^{-1}$

$$v_{\text{ball–Earth}} = v_{\text{ball–train}} + v_{\text{train–Earth}} = (-10) + (+30) = +20 \text{ m s}^{-1}$$

When Bill throws it back, the ball will be moving at:

$\longleftrightarrow$ $- \quad +$

$v_{\text{ball–train}} = +10 \text{ m s}^{-1}$

$v_{\text{train–Earth}} = +30 \text{ m s}^{-1}$

$$v_{\text{ball–Earth}} = v_{\text{ball–train}} + v_{\text{train–Earth}} = (+10) + (+30) = +40 \text{ m s}^{-1}$$

b To determine the force, we need to know the ball's acceleration. The change of velocity, from Angela and Bill's point of view, is:

$\longleftrightarrow$ $- \quad +$

$u_{\text{ball–train}} = -10 \text{ m s}^{-1}$

$v_{\text{ball–train}} = +10 \text{ m s}^{-1}$

$$\Delta v_{\text{ball–train}} = v_{\text{ball–train}} - u_{\text{ball–train}} = (+10) - (-10) = +20 \text{ m s}^{-1}$$

From Clare's point of view the change of velocity was:

$\longleftrightarrow$ $- \quad +$

$u_{\text{ball–Earth}} = +20 \text{ m s}^{-1}$

$v_{\text{ball–Earth}} = +40 \text{ m s}^{-1}$

$$\Delta v_{\text{ball–Earth}} = v_{\text{ball–Earth}} - u_{\text{ball–Earth}} = (+40) - (+20) = +20 \text{ m s}^{-1}$$

Notice that the change of velocity was the same in both frames of reference. This change took 2 s and so in both frames of reference the acceleration was:

$\longleftrightarrow$ $- \quad +$

$\Delta v_{\text{ball}} = +20 \text{ m s}^{-1}$

$\Delta t = 2 \text{ s}$

$$a_{\text{ball}} = \frac{\Delta v_{\text{ball}}}{\Delta t} = \frac{+20}{2} = +10 \text{ m s}^{-2}$$

And the force that Bill exerted on the ball was therefore:

$\longleftrightarrow$ $- \quad +$

$a_{\text{ball}} = +10 \text{ m s}^{-2}$

$m = 0.2 \text{ kg}$

$$F_{\text{Bill on ball}} = ma_{\text{ball}} = (0.2)(+10) = +2 \text{ N}$$

Figure 6.16
Galileo tried to measure the speed of light by timing its travel from his lantern to a friend on a nearby hill and back again.

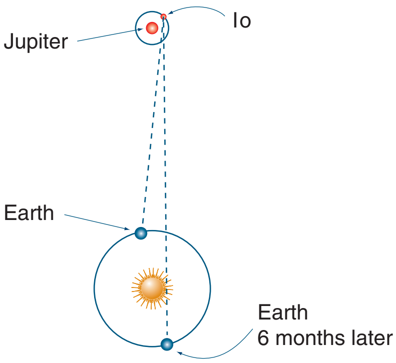


Figure 6.17
Ole Römer determined the time taken for the light from Jupiter's moons to cross the diameter of the Earth's orbit.

Figure 6.18
Michelson used a rotating octagonal mirror to measure the speed of light. In the 1920s, he set up his apparatus at the Mt Wilson Observatory and placed the stationary mirror on a mountain 35 km away.

You were probably not surprised about the result of worked example 6.2A. Indeed, it would be very strange if the force Bill exerted on the ball depended on who was doing the measurement! The point is that in frames of reference that are moving steadily relative to one another, while the velocity is different, any changes of velocity (and, therefore, accelerations and forces) will be just the same. This is one expression of the Galilean–Newtonian principle of relativity and, as we shall see, it was a principle that Einstein felt could not be given up. The other idea essential to Einstein's relativity concerns the nature and speed of light.

The speed of light

The speed of light has fascinated physicists for a very long time. Galileo was one of the first to try to measure it. He sent an assistant, equipped with a lamp and shutter, to the top of a hill several kilometres away. When Galileo uncovered his lamp, the assistant was to uncover his. By timing the interval between uncovering his own lamp and seeing the return signal, Galileo hoped to measure the speed of light. Galileo was disappointed to find that no matter what the distance between him and his assistant, he found the same time interval, which was basically the assistant's reaction time. Light either travelled instantaneously or was much faster than Galileo could measure with this technique. Many others tried similar experiments, but with no more success than Galileo.

It was not until about 1675 that Danish astronomer Ole Römer succeeded in showing that light did have a finite speed. He had decided that light was too fast for Earth-bound experiments and so decided to look for a 'clock' in the heavens with which he could time light over astronomical distances. The moons of Jupiter, which Galileo had discovered 60 years earlier, provided just what he needed. By this time, the periods of the moons had been calculated quite accurately, but a slight anomaly had been noticed. The periods seemed to vary a little. Römer realised that the variation was related to the time of year and suggested that it was due to the longer time taken for the light from the moons to reach the Earth when Earth was on the side of its orbit that was furthest from Jupiter.

After a careful analysis of the data, Römer concluded that the moons seemed to be about 22 minutes 'late' when the Earth was furthest from Jupiter, compared with the times when it was closest. This, he concluded, was the time it took for light to travel across the diameter of the Earth's orbit. Römer was more interested in showing that light did indeed have a finite speed than in an accurate measurement; for one thing, the radius of the Earth's orbit (1 AU) was not accurately known at that time. It was clear from his analysis, however, that light travelled at a finite, but very high, speed.

The first reasonably accurate, Earth-bound measurement of the speed of light was done by Frenchman Louis Fizeau in 1849, but it was not until 1880 that Albert Michelson obtained reliably accurate determinations of the speed. Both Fizeau and Michelson 'chopped' light into pulses and sent them quite some distance, from where they were reflected back again. We will look briefly at Michelson's method.

Michelson used a rotating octagonal mirror to send pulses of light to a mirror some distance away. When the mirror was in the position shown (Figure 6.18), light was reflected from the spotlight to the distant mirror. Clearly only a very brief flash returned from the distant mirror, but if the speed of the rotating mirror was such that on its return the mirror had rotated exactly one-eighth of a turn, the light was reflected back to the observer by the next segment of the mirror. Careful measurements of the speed of the rotating mirror enabled Michelson to establish a very accurate value for the speed of light. He refined his method over the next 40 years, eventually obtaining a result that was within 0.05% of today's value. He was also able to measure the speed of light in various materials such as water, as well as in a vacuum. These measurements confirmed the relationship between refractive index and the speed of light in the medium.

The speed of light is one of the most accurately measured constants in physics and its value has been determined very precisely by many different methods.

The accepted value of the ***SPEED OF LIGHT*** in a vacuum is now:
$c = 2.997\ 924\ 58 \times 10^8$ m s^{-1}

In the early 1800s, Thomas Young, with his famous two-slit experiment, had shown convincingly that light was some type of wave. By the mid-1800s Fizeau had measured the speed at which the wave travelled. The question, however, was what was it that was waving, and what medium carried the waves? Sound waves were clearly pressure waves travelling through air, but light travelled across the vast vacuum of space. These were the questions that concerned a young Cambridge physicist who had studied under Michael Faraday who, significantly for our story, had made the suggestion that light could be some sort of electromagnetic phenomenon. That young physicist was James Clerk Maxwell.

Maxwell's conundrum

In 1864, 250 years after Galileo's major work, Maxwell introduced the second radical idea crucial to our story of relativity. Maxwell was a brilliant mathematician and theoretical physicist. He had taken Faraday's imaginative concept of electric and magnetic fields and worked with them to produce a mathematical description that encompassed all known electromagnetic phenomena. This description can be summed up in four famous equations referred to as Maxwell's equations (Figure 6.19).

A good theoretical physicist tries not only to describe known phenomena mathematically, but to use the mathematics to predict new phenomena. Maxwell did just this. He found that his equations could be used to predict that changing perpendicular electric and magnetic fields could 'self-propagate' through space at a speed given by the fundamental electric and magnetic force constants. This speed turned out to be 3.00×10^8 m s^{-1}, very close to the speed that Fizeau had measured for light a little over 10 years earlier. Maxwell was convinced that this was no coincidence. Maxwell declared that light was an electromagnetic wave, just as predicted by his mentor, Michael Faraday.

Maxwell faced a conundrum, however. His derivation of this speed showed that its value depended only on the two 'constants' for the strength of the electric and magnetic force fields near charges or currents. There was no indication that it should be any different if, for example, the source of the light was in motion. This in itself was not surprising; the speed of sound waves does not depend on the speed of the source either. (The perceived frequency of sound changes if the source is in motion, but not the measured speed.)

Physics file

The value of $c = 2.997\,924\,58 \times 10^8$ m s^{-1} for the speed of light is not just accepted, it is now defined as such. Rather than defining speed in terms of length and time, the speed of light is now the primary standard with the unit of length, the metre, being defined as the distance light travels in a time equal to $\frac{1}{2.997\ 924\ 58 \times 10^8}$ seconds.

This value is consistent with the old standard—the length between two marks on a special bar kept at Sèvres near Paris. In fact, the distance between the marks cannot be measured to this degree of accuracy.

$$\oint \vec{E} \cdot d\vec{A} = q/\epsilon_0$$

$$\oint \vec{B} \cdot d\vec{A} = 0$$

$$\oint \vec{E} \cdot d\vec{s} = -\frac{d\Phi_B}{dt}$$

$$\oint \vec{B} \cdot d\vec{s} = \mu_0 \epsilon_0 \frac{d\Phi_E}{dt} + \mu_0 i$$

Figure 6.19
Maxwell's equations are only to be admired, not learnt! The first describes the electric field around a charge and the second describes the fact that magnetic fields are continuous. The third is Faraday's law of electromagnetic induction and the fourth tells us that magnetic flux is generated by currents or changing electric flux.

Physics file

If light was an electromagnetic wave, were there other types of electromagnetic waves? This was another obvious question physicists asked themselves after the publication of Maxwell's work. Just 7 years after Maxwell's rather early death at 48, Heinrich Hertz demonstrated the existence of lower frequency electromagnetic waves by transmitting an electrical effect from one coil, in which there was an oscillating electric current, to another coil a short distance away. This was the beginning of radio—a very practical outcome from the highly theoretical work of Maxwell!

It was surprising, however, that there appeared to be no allowance for the speed of the 'receiver' of the light. If you run towards a ball thrown at you, the speed of the ball, relative to you, is clearly greater than if you simply stand still. The same is true for sound. If you are moving towards a source of sound, the speed of the sound (the product of the measured frequency and wavelength) will be the sum of the speed of sound in the medium and your own speed. Maxwell's equations, however, simply refused to make any allowance for the motion of the observer of light waves! Whatever the speed of the observer, the measured speed of light should be just 3.00×10^8 m s^{-1}—something quite against all the known laws of physics, and particularly in complete contradiction to the principle of Galilean relativity.

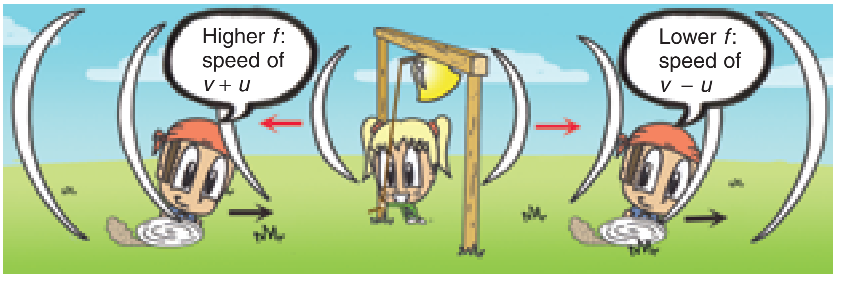


Figure 6.20

As Bill runs towards Angela ringing the bell, the speed of sound will seem higher than normal. As he runs away, it will seem lower. In both cases, the speed is equal to $f\lambda$, where the wavelength will be the same, but in the first scenario, the frequency will seem higher, and in the second, it will seem lower.

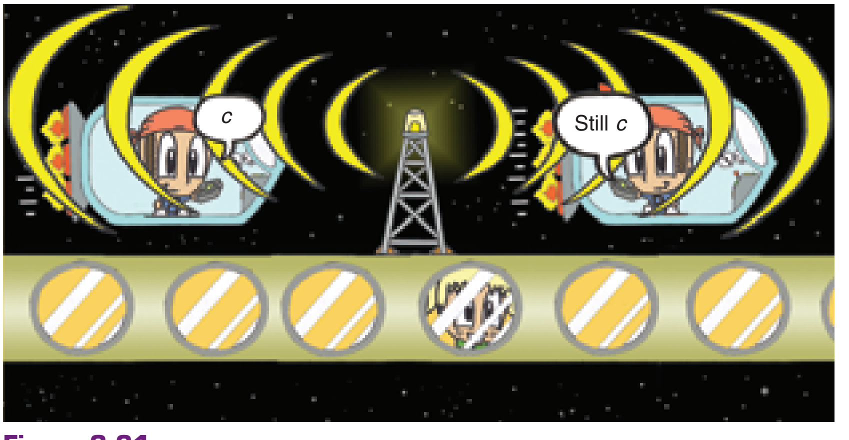


Figure 6.21

This time, as Bill speeds towards Angela's space station beacon light, he measures the apparent speed of light to be *c*. Maxwell's equations say that as Bill moves away from the space station, he will find that the speed is still exactly *c*.

Virtually all physicists, including Maxwell himself, felt that some mistake would be found in his reasoning to rectify this apparent paradox. It was thought that the speed predicted would be the speed in the medium in which light travelled, and the measured speed would have to be adjusted for one's own speed through that medium. This medium, however, was another difficulty! As light travelled through the vacuum of space between the Sun and Earth, clearly the medium was no ordinary material. Physicists gave it the name *aether*, as it was an 'ethereal' substance, if indeed it could be called a substance at all. It was thought, following Maxwell's work, that the aether must be some sort of massless, rigid medium that 'carried' electric and magnetic fields.

The existence of an aether appeared to be a serious blow for the principle of relativity. It seemed that there might be after all a frame of reference attached to space itself. If this was the case, there was the possibility of an absolute zero velocity. What the laws of mechanics had failed to do, Maxwell's laws of electromagnetism had apparently done.

For all practical purposes, this difficulty with the speed of electromagnetic waves was not a problem. No ordinary earthly velocity could possibly approach anything like the speed of light and so any discrepancy in the measured speeds would be quite undetectable. Even the speed of the Earth in its orbit around the Sun was only one ten-thousandth of the speed of light. Physicists are not to be put off by such practicalities, however! How could this idea of electromagnetic waves moving through the aether be tested?

The Michelson–Morley experiment

Presumably, it was thought, the Earth itself must be moving through the aether in its orbit around the Sun. Perhaps the Sun was at rest in the aether? (Remember that this was well before the discovery of galaxies and it was quite reasonable to think of the Sun as the centre of the Universe.) This suggested to American physicist Albert Michelson that it should be possible to measure the speed at which the Earth was moving through the aether by measuring the small changes in the speed of light as the Earth changed its direction of travel. For example, if the light was travelling in the same direction as the Earth, through the aether, the apparent speed should be slower than usual, but if the light was travelling against the Earth's motion, the apparent speed should be faster. The differences would be tiny, less than 0.01%, but Michelson was confident that he could measure them.

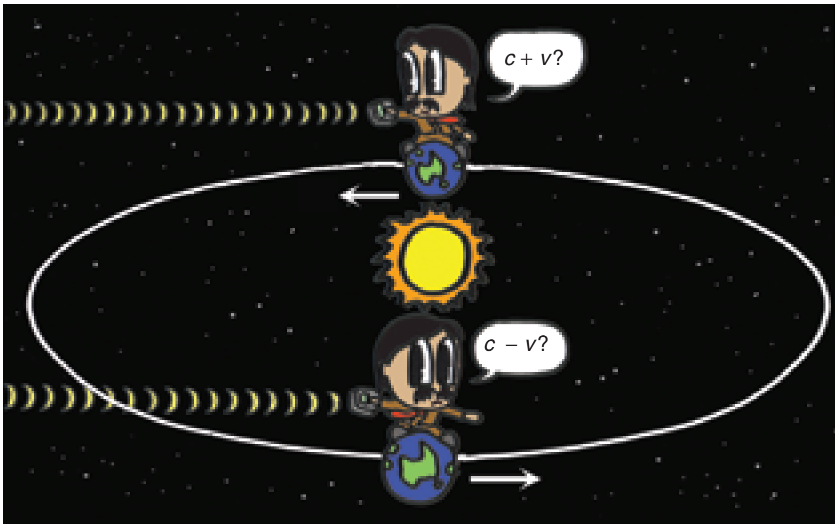


Figure 6.22

The basic principle of the Michelson–Morley experiment. If the aether is fixed relative to the Sun, and the light is travelling (at c relative to the aether) in the same direction as the Earth, the apparent speed should be less than c. Six months later, light travelling in the same direction should appear faster than c. The details were actually a little more complex.

In what is now one of the most famous physics experiments, in the 1880s Michelson and his collaborator Edward Morley set up a device known as an interferometer, which was able to measure the very small differences in the time taken for light to travel in two mutually perpendicular directions. They were able to rotate the whole apparatus and hoped to detect the small difference that should result from the fact that one of the directions was to be the same as that in which the Earth was travelling and the other at right angles. However, they found no difference. Perhaps, then, the Earth at that time was stationary with respect to the aether?

Six months later, however, when the Earth would have to be travelling in the opposite direction relative to the aether, there was still no difference in the measured speeds! Other people performed similar experiments with many different variations, virtually always with the same result. Whatever direction the Earth was moving, it seemed to be at rest in the aether. Some thought that maybe the Earth dragged the aether around with it, but this was shown to be inconsistent with other principles of physics.

While Michelson and Morley's results were consistent with Maxwell's prediction that the speed of light would always appear to be the same for any observer, the apparent absurdity of such a situation led most physicists to believe that some flaw in the theory behind the experiment, or in its implementation, would soon be discovered. One interesting approach was put forward by Dutch physicist H. A. Lorentz. He suggested that the null result of the Michelson–Morley (M–M) experiment could be explained if moving objects contracted very slightly in the direction of their motion. This would mean that the path of light in the M–M experiment, which was in the direction of the Earth's motion, would contract a little, just enough for the light to arrive at the observer at the same time as light from the other (perpendicular) path. He even worked out a formula for the contraction which would give the null result. However, although this was a clever idea, there seemed to be no physical explanation for this contraction. Neither did it offer a satisfactory explanation of the role of the aether or of the inconsistency with the principle of relativity.

While most physicists were looking for the problems with Maxwell's equations or the M–M experiment, there was one young physicist with a particular interest in the nature of light, and a very sharp mind, who was convinced that Maxwell's equations were so elegant and so well founded that they just had to be true. He wondered about the consequences of actually accepting their prediction about the speed of light but at the same time holding on to that other elegant piece of physics, the relativity principle. This physicist's name was Albert Einstein.

Figure 6.23
Einstein as a teenager.

Einstein's crazy idea

Albert Einstein was a daydreamer—a theoretical physicist. When he was just 5 years old, his father gave him a compass. He was fascinated by the fact that it was responding to some invisible field that enveloped the Earth. His curiosity was aroused and, fortunately for physics, he never lost it. In his teens, his daydreaming turned to the question of light. What, he wondered, would it be like to 'ride a beam of light'?

As he studied physics more, Einstein became familiar with Maxwell's work and was fascinated by it. He became convinced that the answer to his question about riding a light beam was that you couldn't. If somehow you could catch up to a light beam, what would those waving fields of electromagnetism look like? The answer was that the waves would appear frozen in time. You would see fixed electric and magnetic fields, changing with position, and apparently coming from nowhere. Nothing like that had ever been seen, and Maxwell's laws suggested that it was not possible. Einstein was sure, for these sorts of reasons, that it must be impossible to travel fast enough to see light slow down and stop. The idea of 'frozen light' simply didn't make sense. Whenever light is 'stopped' by something, a piece of black paper for example, we never see stationary light. All we ever see is the heat from the energy that was carried by the light.

Perhaps it was lucky that in his early twenties Einstein was not actually part of the physics 'establishment'. Partly as a result of his curiosity about the nature of light, he had not really taken his other studies seriously enough to obtain an academic position. As a result, he was working as a patent clerk in the Swiss patent office in Berne. Although this was quite an interesting job, it also left him plenty of time to mull over his ideas about light and electromagnetic waves and their relationship to the Galilean principle of relativity. He and some friends, originally students he had taken on to tutor, would spend hours in the local coffee shop freely exploring ideas that perhaps the academic establishment would have frowned upon.

A characteristic of a good theoretical physicist is that they like tidy things—the messy world of the laboratory is not for them! Einstein was the archetypal theoretician; the only significant experiments he ever did were thought experiments, or '*Gedanken*' as they are called in his first language (German). Many of his Gedanken experiments involved thinking of situations that involved two frames of reference moving with a steady relative velocity, in which the principles of Galilean relativity applied. Newton had referred to these as *inertial frames of reference*, as the law of inertia applied within them, but not within frames that were accelerating. We could think of the smooth quiet train we discussed earlier as a 'Gedanken train'. We don't need to worry about the practicalities of making a real one; we just imagine one!

The elegance of physics

Einstein decided that the elegance of the principle of Galilean relativity was such that it simply had to be true, despite the problems with light. Nature did not appear to have a special frame of reference, and Einstein could see no reason to believe that there was one waiting to be discovered. In other words, there is no such thing as an *absolute velocity*. It is not possible to have a velocity relative to space itself, only to other objects within space. This is equivalent to saying that space itself has no 'centre' or 'edges', no built-in set of *xyz* coordinates, and nothing upon which to attach the mysterious aether.

He expanded the Galilean principle to state that all inertial frames of reference must be equally valid, and that the laws of physics must apply equally in any frame of reference that is moving at a constant velocity. A consequence of this is that there is no physics experiment you can do, completely within a particular frame of reference, to tell that you are moving. In other words, as you speed along in our Gedanken train with the blinds down, you cannot measure your speed, at least not without peeking out the window, or connecting something to the wheels, which are in contact with the outside frame of reference. You can tell if you are accelerating easily enough: just hang a pendulum from the ceiling. However, the pendulum will hang straight down if your velocity is constant, whether you are travelling steadily at 100 km h^{-1} or are stopped in the station.

To understand why Einstein was so sure that the principle of Galilean relativity had to be true, let's perform another Gedanken experiment. Imagine, for a moment, that some physical law *does* vary with an absolute velocity. Maybe the force between two electric charges increases with absolute speed. If that were the case, we could make a little device, let's call it a *veelo*, based on the measurement of the force between two fixed

Figure 6.24

Einstein as a young man was quoted as saying, 'I have no particular talent. I am merely inquisitive.'

Physics file

Newton realised that the principle of relativity applied to any frame of reference that was not accelerating. These he referred to as *inertial frames of reference*. In a non-inertial frame of reference—that is, one which is accelerating—Galileo's principle of inertia (which of course we usually call Newton's first law) would not work. Any object on which there was no net force would appear to accelerate in the opposite direction. In order to stay at rest in the frame of reference, there would appear to be mysterious forces acting. For example, when we turn a corner in a car, we experience a non-inertial frame of reference; there seems to be a mysterious force pushing us to the side of the car. Try riding in a car with your eyes closed. Fairground rides exploit these unusual forces to make us feel rather odd!

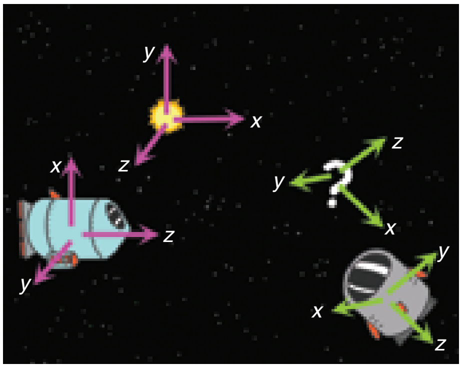


Figure 6.25
We can make measurements relative to stars and spaceships, but there is nothing in 'free space' to attach our coordinates to.

charges, and tell our speed without any reference to the outside world. We could pull down the blinds in our Gedanken train and use our veelo to find the speed. But wait a minute, the Earth is turning on its axis: does our veelo measure the speed at which we are moving down the track or at which we are rotating around with the Earth? Or maybe the orbital speed relative to the Sun? Perhaps it is one of those, but now let us take our veelo out into space a long way from our solar system.

In the distance we see a Norgon spaceship (from the star system Norg) travelling towards us; or is it travelling in the same direction but more slowly than we are? We might discuss our speed in relation to the Sun we left many years ago, but that speed is quite meaningless to the people from Norg in the other craft. So if we use our veelo, what speed could it tell us about? And if the Norgons also had discovered veelos, what speed would theirs tell them about? What would happen if we swapped veelos? This sort of thinking convinced Einstein that the idea of an absolute frame of reference, or absolute velocity, just didn't make sense. The principle of relativity says that there is no point in trying to work out absolute speeds; all we can ever do is discuss relative speeds.

So Einstein decided that the relativity principle could not be abandoned. Whatever the explanation for the strange behaviour of light, it could not be based on a flaw in the principle of Galilean relativity.

Einstein's fascination with the nature of light had led him to a deep understanding of Maxwell's work on the electromagnetic nature of light waves. He was convinced of the elegance, indeed the beauty, of Maxwell's equations and their implications for the speed of light. As we have seen, however, most physicists believed that the constant speed predicted by Maxwell's equations referred to the speed of light relative to the aether. This was a real problem for Einstein. A speed of light fixed in the aether would be in direct conflict with the principle of Galilean relativity which Einstein was reluctant to abandon.

As in any conflict, the resolution is usually found by people who are prepared to look at it in new ways. This was the essence of Einstein's genius. Instead of looking for faults in what appeared to be two perfectly good principles of physics, he decided to see what happened if they were both accepted, despite the apparent contradiction.

So Einstein swept away the problem of the aether, saying that it was simply unnecessary. It had been invented only to be a 'medium' for light waves and no one had found any evidence for its existence. Electromagnetic waves, he said, could apparently travel through empty space without a medium in which to travel. Doing away with the aether did not solve the basic conflict between the absolute speed of light and the principle of relativity, however. How could two observers travelling at *different* speeds see the same light beam travelling at the *same* speed?

The special theory of relativity

The answer, Einstein said, was in the very nature of space and time. In 1905 he sent a paper to the respected physics journal *Annalen der Physik* entitled 'On the electrodynamics of moving bodies'. In this paper he put forward two simple postulates and followed them to their logical conclusion. It was this conclusion that was so astounding.

EINSTEIN'S TWO POSTULATES:

I The principle of relativity states that no law of physics can identify a state of absolute rest. (In other words, no single inertial frame of reference is better than any other.)

II The speed of light will always be the same no matter what the motion of the light source or observer.

Einstein's original postulates are quoted in the Physics in action on page 284. The first postulate is basically that of Newton, but Einstein extended it to include the laws of electromagnetism so elegantly expressed by Maxwell. The second postulate simply takes Maxwell's prediction about the speed of electromagnetic waves at face value. We should add here that we are referring to the speed of light in a vacuum.

These two postulates sound simple enough; the only problem was that, according to early 1900s physics, they were contradictory. If Angela is in her spaceship travelling away from Bill at a speed v, and Bill turns on a laser beam to signal Angela, postulate I would imply that the speed of the laser light as measured by Angela should be $c - v$, c being the speed of light in Bill's frame of reference. (Space travel may not have been achieved in those times, but that didn't stop physicists from having Gedanken spaceships!) This is what we would expect if, for example, we measure the speed of sound as we travel away from its source; the nearer we get to the speed of sound ourselves, the slower the sound appears to be travelling. Postulate II, however, tells us that when Angela measures the speed of Bill's laser light she will find it to be c; that is, 3.00×10^8 m s^{-1}, no more and no less. So at first glance, these two postulates appear to be mutually exclusive.

How can this paradox be resolved? This is where Einstein's brilliant mind had to work doubly hard, but before we consider his answer to this question, we need to turn back the clock to another brilliant mind, that of Isaac Newton in 1686 with the publication of his famous *Principia*. Right at the start of this incredible work, which laid the basis for all physics in the following two centuries and more, he states various assumptions that he makes. He includes this statement:

The following two statements are assumed to be evident and true:

1 Absolute, true and mathematical time, of itself, and from its own nature, flows equably without relation to anything external.

2 Absolute space, in its own nature, without relation to anything external, remains always similar and immovable.

Newton's genius was such that he realised that all his laws were based on these two assumptions; that space and time are how they seem to us: constant, uniform and straight. That is, space is like a big set of *xyz* axes which always remain mutually perpendicular and in which distances can be calculated exactly according to Pythagoras's rule. In this space, time flows on at a constant rate which is the same everywhere. Certainly we may have to adjust our clocks as we fly around the Earth, but one second here is the same as one second there, and one second on the ground is the same as one second as we fly. Similarly, we expect a metre rule to be the same length whether it is in our classroom at school or flying around the Earth in the International Space Station.

Physics file

It should be pointed out here that even if the speed of light is the same for all observers, motion towards the source of light will result in what is called a Doppler shift, the increase of frequency due to the fact that we are encountering more waves every second. (Or the reverse if we are moving away.) This effect is familiar in sound when the higher and then lower pitch of the siren as an ambulance approaches and then recedes from us (see Figure 6.20). The well-known redshift in the light of distant galaxies is a Doppler shift in the frequency of the light due to the high velocity of those galaxies moving away from us. You may well ask, then, if the frequency of the light has decreased, doesn't this mean (as $v = f\lambda$) that the speed will have decreased also? Strangely, the answer is no, we still measure the same speed. This implies that λ has increased, but this is quite different from the behaviour of 'normal' waves! This is a hint as to what relativity is all about: space (and therefore lengths) can behave in very strange ways.

Einstein realised that the assumptions Newton made, and everyone else since, may not in fact be valid, at least not on scales involving huge distances and speeds approaching that of light. The only way in which postulates I and II can both be true, Einstein said, is if space and time are, in fact, not 'absolute'. But what on Earth does that mean?

Physics in action — Einstein's 'Electrodynamics of moving bodies'

Einstein's 1905 paper on relativity was the third of five papers he published in that year. Each of the papers was a remarkable achievement in its own right. It was actually the fifth, on the photoelectric effect, that eventually resulted in his award of the Nobel Prize in Physics for 1921. Interestingly he did not receive the Nobel Prize for his work on relativity. Its real significance was still not universally recognised nearly 20 years after the first publication.

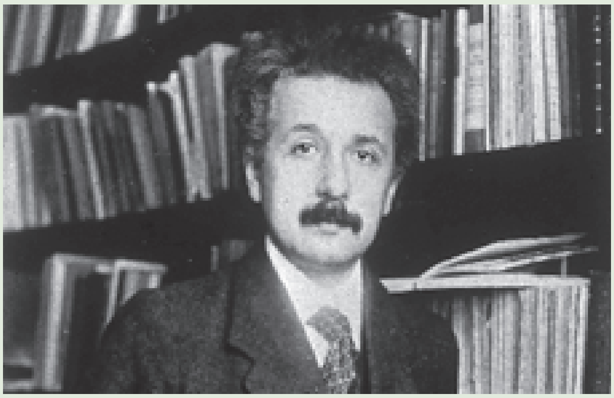


Figure 6.26
Einstein in 1905.

The papers were published in the prestigious German physics journal *Annalen der Physik*. The title of the third, on relativity, and published on 30 June 1905, was 'Zur Elektrodynamik bewegter Körper', or 'On the electrodynamics of moving bodies'. You can see immediately from the title that it was a paper heavily influenced by Maxwell's work on electromagnetism. Einstein introduced the paper with a discussion of the fact that whether one moves a magnet near a wire, or a wire near a magnet, the current induced in the wire is the same, although the theory used to deduce the current is different in each case. It is only the *relative* motion that is important. He points out that no experiments in physics—in mechanics, optics or electromagnetism—can identify any form of *absolute* motion. Only relative motion can be detected or measured. He goes on:

> Examples of this sort, together with the unsuccessful attempts to detect a motion of the earth relative to the 'light medium', lead to the conjecture that not only the phenomena of mechanics, but also those of electrodynamics, have no properties that correspond to the concept of absolute rest. Rather, the same laws of electrodynamics and optics will be valid for all coordinate systems in which the equations of mechanics hold, as has already been shown for quantities of the first order. We shall raise this conjecture (whose content will hereafter be called 'the principle of relativity') to the status of a postulate and shall also introduce another postulate, which is only seemingly incompatible with it, namely that light always propagates in empty space with a definite velocity ***V*** that is independent of the state of motion of the emitting body. These two postulates suffice for the attainment of a simple and consistent electrodynamics of moving bodies based on Maxwell's theory for bodies at rest. The introduction of a 'light ether' will prove to be superfluous, inasmuch as the view to be developed here will not require a 'space at absolute rest' endowed with special properties, nor assign a velocity vector to a point of empty space where electromagnetic processes are taking place.
>
> Translation from *Einstein's Miraculous Year*, John Stachel, Ed., Princeton University Press, 1998

The rest of the paper becomes more mathematical and not for the faint hearted! In it he puts forward the main concepts of 'special relativity', including the relativity of lengths and times along with the mathematical transformations involved, a section on the addition of velocities—which shows that no two velocities can add to more than the speed of light—and more. Although the famous $E = mc^2$ does not appear in this paper, it followed later from considerations introduced in it.

On 2D, 3D and 4D worlds

How far is Perth from Sydney? Measured on an atlas, it is about 3400 km in a straight line. But that is how far the crow (or, more likely, aeroplane) has to fly. It is 'really' about 3350 km in a genuine straight line, which would take us through the Earth about 260 km under the Great Australian Bight. This example can help us to get a feel for what Einstein means. We know what we mean when we say Sydney is 3400 km from Perth. We know the line is actually curved, but when we think of lines on maps we think in two dimensions, not three. It is when we move from 2D to 3D thinking that we realise that 'real' distances will be a little different. Now we believe we live in a 3D world, but what if it is really 4D? What would that do to our 'straight' lines?

Our friends Angela and Bill are not convinced that the Earth is round; after all, it looks pretty flat. They decide to do an experiment to check. Angela heads north and Bill heads east and they vow to keep walking exactly in a straight line. If the world really is flat, they know they will never meet again. However, if the world is round, they will meet again halfway around the world. (Remember they are Gedanken people and things like mountains, oceans, storms and ice don't bother them!) In our everyday experience, we can treat the world as 2D; it is only on very long journeys that we need to take into account its 3D nature. Could it be that on extremely long journeys, at very great speeds, we need to take into account its 4D nature?

We have no experience of a 4D world, but that is not to say that it doesn't exist. An ant probably never 'thinks' in terms of a 3D world. It has no need to; two dimensions are fine for all its needs. The occasional strange experience it has, of a human foot suddenly appearing from nowhere, is just that—a strange experience. We could expect that if and when we experience a 4D world, we may well have strange experiences! The interesting thing is, however, that just as Angela and Bill could check to see if their world was 2D or 3D, perhaps we could also check to see if our 3D world is really 4D. Well, we have checked and guess what? We *do* live in a 4D world of *spacetime*! But we are getting ahead of ourselves; let's return to Einstein.

Figure 6.27
Normally we would think of a straight line from Sydney to Perth as being 'as the crow flies', a distance of about 3400 km. An actual straight line would pass through the Earth and be about 50 km shorter.

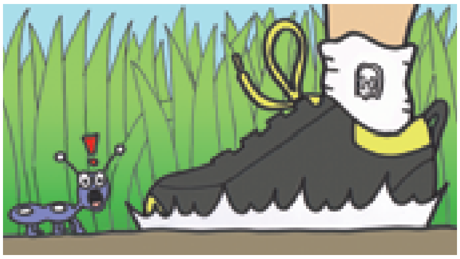


Figure 6.28
When the foot appears in the ant's 2D world, the ant has no idea where it came from.

Einstein's Gedanken train

To illustrate the consequences of accepting the two postulates he put forward, Einstein discussed a simple thought experiment. It involves a train, moving at a constant velocity. Angela and Bill have boarded Einstein's train to help us with our experiments and Clare is still outside on the platform. This train has a flashing light bulb set right in the centre of the carriage. Angela and Bill are watching the flashes of light as they reach the front wall and back walls of the carriage, respectively. (Being Gedanken people they have eyes in the back of their heads and very fast reflexes.) They are not surprised to find that the flashes reach the front and back walls at the same time.

Outside, Clare is watching the same flashes of light. Einstein's interest was in when Clare saw the flashes hit the end walls. Now we realise that the light has to reflect from the walls and then travel to Clare before she can 'see' it hit the wall, but we are going to make the simplifying assumption that she makes the appropriate calculations to determine when the light 'actually' hit the walls.

Physics file

In discussing relativity, there are often situations where we want to know when light reaches a certain point, like the walls of the train. We (or our intrepid train travellers) only 'see' the light reach a point after it has reflected from that point and returned to our eyes. This extra time is called the *look-back time*. We need to calculate the look-back time and subtract it from the observed time to find the actual time to reach the wall. To avoid confusion, we will assume that our observers always do this and only quote the actual time.

To appreciate Einstein's ideas, we need to contrast them with what we would normally expect. Consider the earlier example in which Angela and Bill were throwing balls back and forth in the train. It is important to appreciate that while our outside observer (Clare) sees the various velocities involved differently, the times at which various events occur must be the same.

If we had discussed a pulse of sound waves travelling from the centre of the train, we would find exactly the same result: Clare always agrees with Angela and Bill that the time taken for balls, or sound waves, to reach the end walls is the same. Indeed, we would be very surprised if that were not the case! Normally we would not even bother to work this sort of problem from the fixed frame of reference, so confident are we of the principle of Galilean relativity. But what about light?

Practical activity

36 Frames of reference

Einstein's second postulate tells us that all observers see light travel at the same speed. Angela, Bill and Clare will all see the light travelling at 3×10^8 m s^{-1}, we do not add or subtract the speed of the train. The fact that adding the speed of the train to the speed of light would make no practical difference is not of importance to us; Gedanken experiments are allowed to be as accurate as we like!

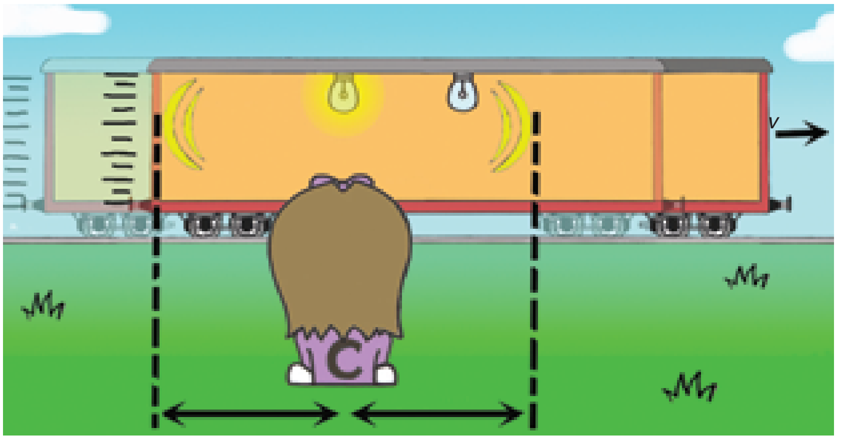


Figure 6.29
Clare sees the flash hit the back end before the front end.

Physics file

What if our hardy observers were riding an open tray truck and shouting into the wind? This time the sound would travel through air at rest in Clare's frame. You might like to confirm that although the shouts would have reached the back end of the carriage first (as it was moving towards the source of the sound), both sets of observers—Angela and Bill in the moving air, and Clare in still air—would have agreed on the time difference. The strange thing about light is that the observers *do not* agree on the times!

If Clare sees the light travelling at the same speed in the forward and back direction, she will see the light hit the back wall first. This is because that wall is moving towards the light, whereas the front wall is moving away from the light and so the light will take longer to catch up to it. Now this is all quite against the principles of Newtonian physics. Angela and Bill saw the light flashes reach the ends of the carriage at the *same* time, Clare saw them reach the walls at *different* times. The idea that two events that are simultaneous for one set of observers are not simultaneous for another is quite outrageous! It is the equivalent of one football umpire seeing a goal kicked before the final siren but another who, because he was running at the time, saw it kicked after the siren (given that they were the same distance from the siren). Remember that Clare did not see the flashes at different times because of some sort of look-back time delay—that has been taken into account.

In the situation with the balls, when we moved from Angela and Bill's frame of reference to Clare's, we knew we had to add the velocity of the train to the velocity of the balls. Had we been discussing the time for simultaneous shouts from Angela and Bill to reach the front and the rear of the carriage, we would again need to add the velocity of the train to the velocity of the sound when we moved from their frame to Clare's. This is because the air in which the sound was travelling was moving with the train. The situation would have been the same as for the ball game and so we would again find that from within both frames the shouts reached the ends of the carriage at the same time.

Simultaneity and spacetime

Practical activity

39 A non-simultaneous simulation

The big difference between the situation for light, and that for balls or sound, is the strange notion that both sets of observers see the speed of light as the same—exactly the same. Whether the carriage is open to the air or closed to the air, the speed of sound in Angela and Bill's frame will always be different from that in Clare's frame by just exactly the speed of the train. For light, however, there is no difference. As a result, events that were simultaneous for one set of observers were not for the others—a very puzzling state of affairs! This is often referred to as a lack of *simultaneity*, simultaneity meaning that we would normally expect that if one set of observers see two events happen simultaneously, we would expect any others to also see them simultaneously.

It needs to be pointed out that our Gedanken people have perfectly good measuring equipment and are not fraught with the usual problems of errors, both systematic and otherwise, in experimental work. If we mere mortals were to attempt these experiments, we would have no hope of detecting the lack of simultaneity as the differences in time we would be trying to measure would be around a millionth of a microsecond, well beyond the capacity of even the best stopwatches! However, while these experiments are purely hypothetical, other experiments based on these ideas are well within the capacity of modern experimental physics and in all cases they confirm Einstein's ideas to a high degree of accuracy—as we shall see later.

How is it possible that two events that were simultaneous to one set of observers were not simultaneous to another? Einstein said that the only reasonable explanation for this is that *time itself* is behaving strangely. The amount of time that has elapsed in one frame of reference is not the same as that which has elapsed in another. In our example, Angela and Bill saw the light flashes that went forward and back take the same time to reach the walls, but in Clare's frame the times were different. Because *time* (which has one dimension) seems to depend on the frame of reference in which it is measured, and a frame of reference is just a way of defining three-dimensional *space*, clearly time and space are somehow interrelated. We now call that four-dimensional relationship *spacetime*. Special relativity is all about spacetime.

Figure 6.30
The famous clock tower in Berne, Switzerland, where Einstein did a lot of thinking about time.

This was a profound shock to the physicists of Einstein's time. Many of them refused to believe that time was not the constant and unchanging quantity that it had always been assumed to be, and certainly always seemed to be. And to think that it might 'flow' at a different rate in a moving frame of reference was too mind-boggling for words. That could mean that if we went for a train trip, our clocks would go slow and we should come back younger than those who stayed behind. Exactly, said Einstein! Well, yes, but younger by something much less than a microsecond—rather difficult to notice. Probably because of the tiny differences involved and the highly abstract nature of the work, many physicists simply scratched their heads and got on with their work. So what, they said, how could it ever have any practical results? How wrong they were.

Consequences of special relativity

There are three consequences of Einstein's special theory of relativity. We will not be discussing these consequences quantitatively by developing equations and then using those equations to calculate the effects; we will limit ourselves to a qualitative understanding of the phenomena.

Time dilation

If we agree with Einstein that the speed of light is constant for each observer, regardless of their frame of reference, then we can compare the period of time for a light event in two frames of reference. This time we will send our travellers, Angela and Bill, on a journey in a Gedanken space ship, while Clare observes from a space station. By observing a photon travelling from the light bulb in our space ship to the floor we can determine the time taken for the event by dividing the distance travelled by the speed of light. Angela and Bill see the light travelling the height of the space ship at the speed of light, so:

$$\Delta t_{\text{Angela and Bill}} = \frac{d}{c}$$

Practical activity

37 Time dilation

At the same time Clare, who is watching from the space station, sees the same event; the photon travels from the light bulb to the floor at the speed of light, but during that time the space ship has moved forward some distance determined by the velocity of the space ship and the period of time for the event. So Clare sees the photon travel at an angle to the ground.

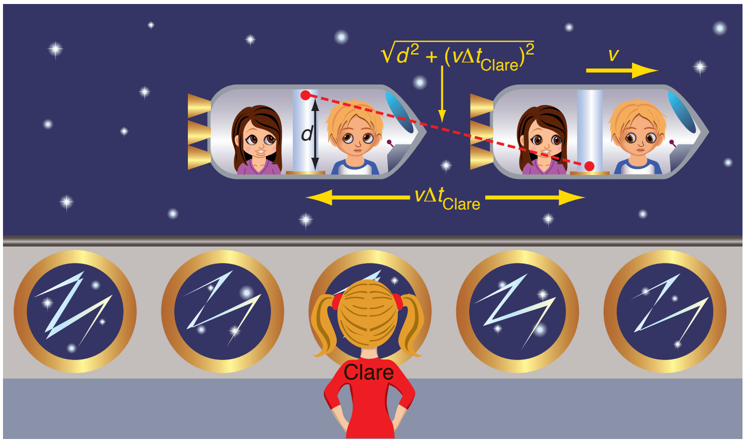


Figure 6.31

A photon event as seen by Clare. The photon has to travel a greater distance at the same speed of light, therefore the event takes a longer time.

The time that the photon takes to travel from the bulb to the floor is:

$$\Delta t_{\text{Clare}} = \frac{\sqrt{d^2 + (v\Delta t_{\text{Clare}})}}{c}$$

The extra distance that the photon must travel in Clare's frame of reference means that the same event takes a longer period of time to occur. In fact every event observed in Angela and Bill's frame of reference by Clare will occur over a longer period of time, the time taken for the second hand on Angela's watch to move will take longer, the time it takes for Bill to draw a breath will take longer. It is important to understand that this is not just an effect like slowing down a movie, it is real for Clare.

It is also important to understand that for Angela her watch works normally and Bill's breathing is at a normal rate too. As Angela and Bill look back to see Clare, they notice that her watch is moving slowly and her breathing is slower too, as their motion is relative. Clare appears to be moving as fast from them as they are moving away from Clare.

The faster the space ship travels, the longer an event takes to occur, when viewed from a stationary frame of reference. If Angela and Bill were riding a photon, how long would an event take when travelling at the maximum speed possible—the speed of light? The answer is—forever! They would appear to Clare to be frozen in time, for Angela and Bill everything else would also appear frozen in time, and it would appear to take no time to travel from one end of the Universe to the other.

TIME DILATION is when an event is timed in an inertial (constant non-zero) frame of reference by an observer in another inertial (stationary) frame of reference, and it will occur over a longer period of time than it does for the same event timed by an observer in the moving frame of reference.

Length contraction

In the previous section, we used a photon travelling from the top of the space ship to the floor to compare the period of time for an event, it was deliberately set up in the spaceship so that this light path, of distance d, was perpendicular to the velocity. The reason for this was that we could expect distances in this perpendicular direction to be unaffected, but this may not be the case for distances in the direction of travel. Indeed, Einstein showed that while perpendicular distances are unaffected, *relative motion affects lengths in the direction of travel.*

Einstein showed that as we observe a fast-moving object, its length, or the distance it travels in a certain time, will appear shorter than the distance we see when the object is stationary in our frame. It will be shorter by the same factor by which time appears to be extended and so Clare would see the length of the space ship contracted when compared to the length as measured by Angela and Bill.

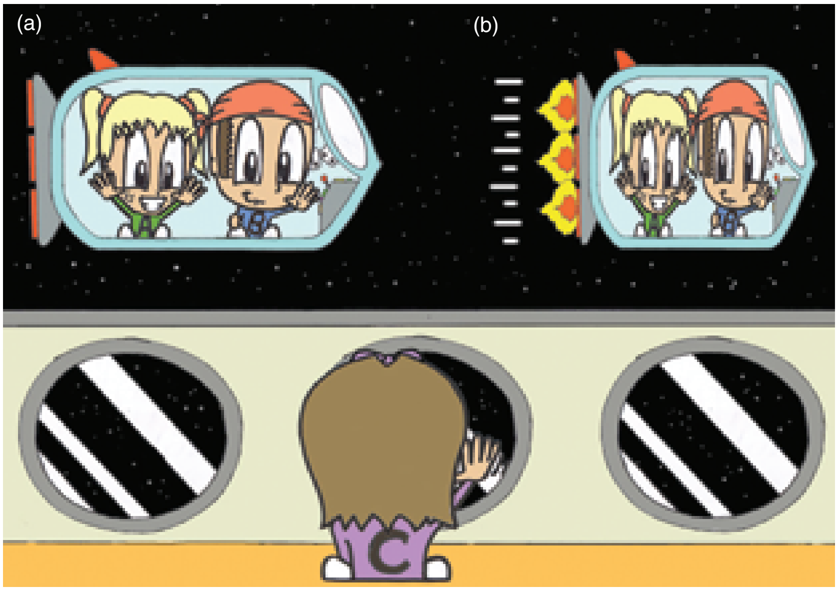


Figure 6.32
(a) Clare watches as Angela and Bill prepare for blast off. (b) As Angela and Bill speed past, Clare sees their rocket ship foreshortened, but the other dimensions remain the same.

Physics file

Einstein said that at the speed of light distances shrink to zero and time stops. No ordinary matter can reach c, but light always travels at c. Strange though it may seem, for light there is no time. It appears in one place and disappears in another, having got there in no time (in its own frame of reference, not ours!). When we stay still, we travel through spacetime in the time dimension only. Light does the opposite: all its spacetime travel is through space and none through time.

Practical activity

38 The Lorentz factor

Remember that length contraction occurs only in the direction of travel, not in any perpendicular direction. To Clare, Angela and Bill's spaceship will appear foreshortened, but its width and height will remain unaltered.

These two simple notions—length contraction and time dilation—are at the core of the special theory of relativity. Together they tell us about spacetime, the four-dimensional world that we inhabit.

A number of significant physicists *did* recognise the importance of Einstein's work. They realised that it would eventually shake the foundations of our ideas about the nature of our world. They were right.

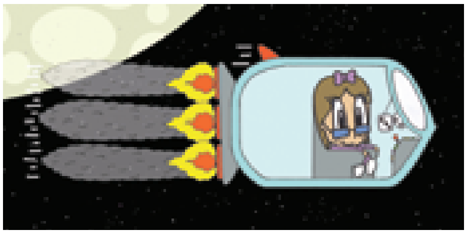


Figure 6.33
This rocket ship is doing 0.99*c* and accelerating, so why can't it accelerate to 1.00*c*?

Relativistic mass

Calculations involving time and distance get a bit rubbery when we consider relativistic effects like time dilation and length contraction. It would not be surprising, therefore, to hear that momentum also displays relativistic effects.

If a rocket ship is travelling at 99% of *c*, why can't it simply turn on its rocket motor and accelerate up to 100%, or more, of *c*? A full answer to this question was not given in Einstein's original 1905 paper on relativity. Some years later he showed that as the speed of a rocket ship approaches *c*, its momentum increases as we might expect, but this is not reflected in a corresponding increase in speed.

Practical activity
50 Distances by parallax measurement

While the force of the rocket engine increases the momentum, near the speed of light the increase in momentum results in smaller and smaller increases in speed. So no amount of impulse can accelerate the rocket ship to the speed of light! One interpretation of this strange behaviour of momentum is that it appears that the inertial mass of the rocket ship is increasing.

Although Einstein himself preferred to stay with relativistic momentum, it is helpful to think in terms of the mass of the rocket ship increasing as its speed approaches *c*. Hence, in terms of Newton's law, we can simply say that as the speed increases, the force required to produce a given acceleration increases towards infinity because the mass approaches infinity. Thus, no matter how much force is applied, or work done, a rocket ship, or anything else, can never reach the speed of light. In fact only things without mass can travel at the speed of light, like photons.

Einstein's famous equation

All physics students know that as the momentum of an object increases, so does the energy. The classical relationship between the two can be written as:

$$E_k = \tfrac{1}{2}mv^2 = \tfrac{1}{2}\rho v$$

We can see, then, that as we approach relativistic speeds the kinetic energy, as well as the momentum, will increase towards infinity. Einstein showed, however, that the classical expression for kinetic energy was not correct at high speeds. The mathematics involved is a little beyond us at this point but Einstein, working from the expression for relativistic momentum and the usual assumptions about work, forces and energy, was able to show that the kinetic energy of an object was given by the expression:

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

where E_k is the kinetic energy of an object corrected for relativistic effects (J), γ is the Lorentz factor, which is the factor that converts rest mass to relativistic mass, m_o is the mass of an object when it is at rest in our frame of reference (kg), and c is the speed of light (3.00×10^8 m s^{-1})

Einstein's expression can be rewritten as:

$$E_k = \gamma m_o c^2 - m_o c^2$$

which in turn can be rearranged as:

$$\gamma m_o c^2 = E_k + m_o c^2$$

Einstein interpreted this expression as being an expression for the *total energy,* of the object. The right-hand side appeared to imply that there were two parts to the total energy: the kinetic energy E_k and another term which only depended on the rest mass, m_o (c being constant). The second term, $m_o c^2$, he referred to as the *rest energy* of the object (as it does not depend on the speed). This appeared to imply that somehow there was energy associated with mass! An astounding proposition to a classical physicist, but as we have seen, in relativity, mass seems to increase as we add kinetic energy to an object.

$$E_{total} = \gamma m_o c^2 = E_k + m_o c^2$$

Now as γm_o is the relativistic mass, m, the total energy can simply be written as:

$$E_{total} = mc^2$$

You will have seen this equation before!

TOTAL ENERGY is given by:

$$E_{total} = E_k + m_o c^2$$
$$= mc^2$$

where m is the relativistic mass.

But what, wondered Einstein, is the significance of the rest energy, the $m_o c^2$ term? As this energy depends directly on the mass of the object, it appears to be energy associated with mass, and so *mass*, he realised, was a form of *energy*! Because the term c^2 is very large (9.00×10^{16} m^2 s^{-2}) clearly the energy associated with just a very small amount of mass is huge. (It is important to realise that c^2 here is acting as a conversion factor, nothing is moving at the speed of light in this situation.) We now understand that this equation tells us that mass and energy are totally interrelated. In a sense, we can say that mass has energy, and energy has mass.

Even a cup of hot water is ever so slightly heavier than the same cup when cold. The difference is far less than we could possibly measure, but it is real nonetheless. It is easy enough to see why: the heat means the molecules have more kinetic energy, and as the total energy is the sum of the rest and kinetic energy, the total energy, mc^2, is greater. This is reflected in the fact that m (relativistic mass) is greater. So a hotter object is heavier! However, because $m_o c^2$ is so much greater than E_k, the increase in mass is not noticeable.

Figure 6.34
In a nuclear bomb a few grams of mass are lost with the energy as the uranium undergoes fission, releasing the equivalent of hundreds of gigajoules (10^{12} J) of energy. Millions of tonnes of TNT (chemical) explosive would be required to produce this much energy.

Physics file

Although you often read that mass is 'converted' to energy, this is not really true. It is *not* that the mass has been converted to heat and destroyed; the mass is still with the energy, wherever it has gone. Another way of picturing it is that when hydrogen and oxygen atoms combine to form water molecules, the mass of the water molecules is a little less than that of the hydrogen and oxygen atoms because the separate atoms have potential energy that the water doesn't (after the heat has dispersed). This extra energy accounts for the extra mass. The potential energy has mass (given by $m = E/c^2$)—not very much, but real all the same.

Nuclear reactions involve vastly more energy (per atom) than chemical ones. When a uranium atom splits into two fission fragments, about 200 million electronvolts of energy is released. By comparison, most chemical reactions involve just a few electronvolts. In this case, it is possible to find the mass of the uranium atom and fission fragments accurately enough to determine the difference. Sure enough, the difference agrees exactly with the prediction of Einstein's famous equation. Likewise, nuclear fusion reactions deep inside the Sun release the huge amounts of energy that stream from the Sun, resulting in a loss of about 4 million tonnes of mass every second.

6.2 SUMMARY Einstein's special theory of relativity

- Galileo's principle of inertia implies that there is nothing special about a velocity of zero.
- Galilean relativity states that the laws of motion cannot determine an absolute velocity. All velocities are relative.
- Maxwell's electromagnetic equations were interpreted to suggest that an absolute frame of reference (the aether) existed in which light always travelled at 3.00×10^8 m s^{-1}.
- Experiments such as Michelson and Morley's failed to detect the Earth's motion through the aether.
- Einstein decided that Galileo's principle of relativity was so elegant it simply had to be true and he was also convinced that Maxwell's electromagnetic equations, and their predictions, were sound.
- His two postulates of special relativity can be abbreviated to:
  - I No law of physics can identify a state of absolute rest.
  - II The speed of light is the same to all observers.
- Einstein realised that accepting both of these postulates implied that space and time were not absolute and independent, but were related in some way.
- Two events that are simultaneous in one frame of reference are not necessarily simultaneous in another.
- This implies that time measured in different frames of reference might not be the same. Time and space are related in a four-dimensional universe of spacetime.
- The pulses in a light clock in a moving frame of reference have to travel further when observed from a stationary frame.
- Because of the constancy of the speed of light, this effectively means that time appears to have slowed in the moving frame.
- The special theory of relativity says that time and space are related. Motion affects space in the direction of travel.
- A moving object will appear shorter, or appear to travel less distance.
- Relativistic momentum implies that as more impulse is added, the mass seems to increase towards infinity as the speed gets closer, but never equal, to c.
- The rest energy is energy associated with the rest mass of the object $E_{rest} = m_0c^2$ and so mass and energy are seen as different forms of the same thing. If a mass Δm is lost, energy $E = \Delta mc^2$ will appear.

6.2 Questions

In the following questions, use 6.37×10^6 m as the Earth's radius and 1.50×10^{11} m as the radius of the Earth's orbit around the Sun.

1 What was the key difference between Aristotle's and Galileo's ideas on motion?

2 a How fast is someone on the Earth's equator moving relative to a person at the south pole?

b How fast are these two people moving relative to a spaceship at rest in the Sun's frame of reference?

c Does the person at the equator have any acceleration? What about the person at the south pole?

3 Römer actually determined that the light from Jupiter's moons was about 22 minutes later in June than in December. What does this value suggest for the value for the speed of light? How close is that to the modern value?

4 In a moving train with the blinds down, it is possible to find one's speed by using a GPS unit. Why does this not violate the principle of relativity?

5 If the speed of sound in air is 346 m s^{-1}, at what speed would the sound from a fire truck siren appear to be travelling in the following situations?
  a You are driving towards the stationary fire truck at 30.0 m s^{-1}.
  b You are driving away from the stationary truck at 40.0 m s^{-1}.
  c You are stationary and the fire truck is heading towards you at 20.0 m s^{-1}.
  d You are driving at 30 m s^{-1} and about to overtake the fire truck, which is travelling at 20.0 m s^{-1} in the same direction.

6 Which of the following are reasonably good inertial frames of reference?
  A An aircraft in steady flight
  B An aircraft taking off
  C A car turning a corner
  D A car driving up a hill of constant slope at a steady velocity

7 a If we were enclosed in a small windowless room which was mounted on a (very smooth) merry-go-round, how could we know if the merry-go-round was rotating?
  b Why doesn't this merry-go-round experiment violate Galileo's principle of relativity?

8 Angela is at the front end of a train carriage moving at 10.0 m s^{-1}. She throws a ball back to Bill, who is 5.00 m away at the other end of the carriage. Bill catches it 0.200 s after it was thrown. Clare is watching all this from the side of the track.
  a At what velocity does Clare see the thrown ball travelling?
  b How far, in Clare's frame of reference, did the ball move while in flight?
  c How long was it in flight in Clare's frame of reference?

9 Imagine that the speed of light has suddenly slowed down to only 50.0 m s^{-1} and this time Angela (still at the front of the 5.00 m train moving at 10.0 m s^{-1}) sends a flash of light towards Bill. (Ignore the look-back time effects for these questions.)
  a How fast was the light travelling in Bill's frame of reference?
  b How fast was the light travelling in Clare's frame of reference?

10 Which (one or more) of these statements would Einstein agree with?
  A The principle of (Galilean) relativity was too elegant to abandon.
  B Maxwell's prediction that the speed of light would always be the same was an error.
  C Two events which were seen to be simultaneous by one observer must also be seen as simultaneous by any other observer.
  D The speed of light depends only on the speed of the light source, not the observer.

6.3 To the stars

Astrophysics is concerned with some of the biggest questions we can ask: What is the Universe? How and why did the conditions for life to evolve occur? The early astronomers thought the stars all lay in the 'realm of the gods', just beyond the planets that were circling the Earth—which was stationary, at the centre of the Universe. Galileo realised both that the Earth circled the Sun and that the stars must be much further away than the planets, or they would show apparent movement as the Earth revolved around the Sun. By Newton's time it was realised that the stars were possibly other 'suns' and therefore must be huge distances away or they would appear much brighter. Newton calculated that if Sirius, the brightest star in the sky, was about the same *intrinsic* (*actual*) *brightness* as the Sun, then it must be about a million times further away. Many people, however, refused to believe that stars could be suns. For thousands of years it was believed that they were 'heavenly' objects quite unlike anything in the solar system, and located just beyond the orbit of Saturn. Distances such as Galileo and Newton were suggesting were simply incomprehensible. So where are the stars, and what are they?

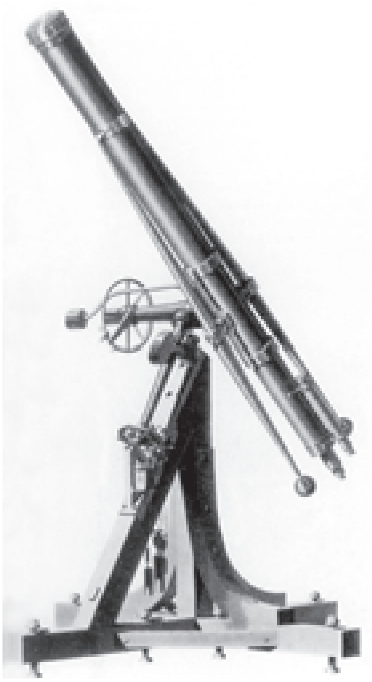


Figure 6.35

Wilhelm Struve used the 'Great Dorpat Refractor' to look for double stars. By 1837 he had found over 3000.

Far far away

The only way the astronomers who followed Galileo could measure the distance to the stars was to look for *parallax* in their positions as the Earth moved around the Sun. The idea is shown in Figure 6.36. As the Earth moves around its orbit the closer star should appear to move in relation to those further away. The same idea can be seen if you simply look out the window and move your head from side to side: the window frame moves 'against' you while the background moves 'with' you.

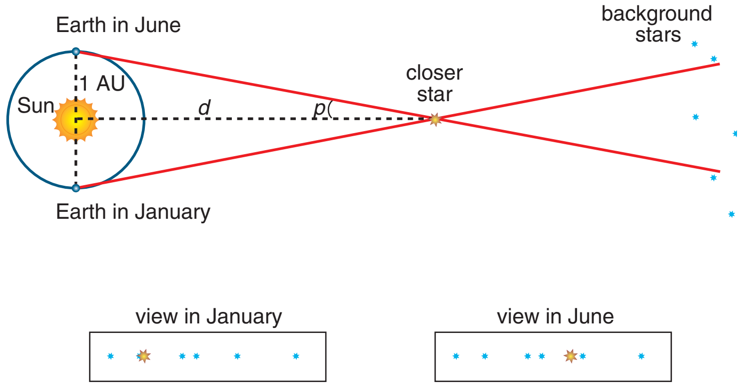


Figure 6.36

The parallax movement of a close star relative to the background 'fixed' stars. The parallax angle p is half the total apparent shift in angle.

Various 18th century astronomers attempted to measure the annual parallax of some bright stars (which they thought more likely to be closer). In 1729 James Bradley announced that if there was any parallax in the positions of stars it was less than 1 arc second. He calculated that they must be at least 400 000 times further away than the Sun (i.e. 400 000 AU). Along with Newton's calculation for Sirius (1 million AU), this meant that attempting to measure the distances to stars seemed futile and so there were few attempts for the rest of the 18th century.

By the beginning of the 1800s instrument makers had developed special telescopes capable of measuring angles of only fractions of an arc second and so attempts were renewed to look for stars showing parallax. By this time it had been realised that some stars move, relative to the background stars, at a steady rate. This was not due to parallax; the star was actually moving (very slowly!) through the sky. For example, the star 61 Cygni had been observed to move 5 arcsec every year (at which rate it would take over 700 years to move 1°). This motion is known as *proper motion* and is very interesting to astronomers as it tells them about the motion of the stars within the galaxy. The early 19th century astronomers realised that stars with a large proper motion were more likely to be closer to us than those with none. Clearly it was also likely that brighter stars were closer. There was one other indication they used to select stars as possible candidates for parallax measurement.

The development of better telescopes had led to the discovery that many stars were actually *binary stars*. That is, they appeared to be two stars in orbit around each other. The further apart the two stars appeared to be (in relation to the time they took to orbit each other), the closer to the Earth the stars were thought to be. So a systematic examination was made of stars that fulfilled at least two of these three criteria. In 1835 German-born Wilhelm Struve, working in Russia, decided to look at Vega, a single star, but a bright one with a large proper motion. He looked for parallax against nearby stars, just as Galileo had suggested many years before. Two years later he announced that Vega showed a parallax of just one-eighth of an arc second (quite close to the modern value)—a tiny value, but he was confident that it was real parallax. It suggested that Vega was a huge 1 600 000 AU from the Earth. Soon other 'close' stars were found. In 1838 Thomas Henderson, who had been working from the Cape of Good Hope, found that Alpha Centauri, a bright binary with a high proper motion, had a parallax of about 1 arcsec, putting it at about 200 000 AU. Alpha Centauri is the 'Pointer' furthest from the Southern Cross. In fact, it is a triple system with a third faint star called Proxima Centauri, for obvious reasons, being slightly closer. The modern value for the parallax of Proxima Centauri is 0.77 arcsec—making it the closest star to our Sun at a distance of 270 000 AU.

This method of measuring the distance to stars is known as *stellar parallax*. It leads to a natural unit for the distance of stars, the **parsec**. The *parsec* (abbreviation pc) is the distance to a star that shows a *par*allax angle of 1 arc *sec*ond. Clearly a star which has a parallax of only 0.5 arcsec will be twice as far away as one with 1 arcsec of parallax, and so on. Thus the distance in parsec is simply the reciprocal of the parallax angle.

In the 20th century, although the parallax of many thousands of stars was measured, most stars were too far away to show any measurable parallax. As a result, astronomers looked for indirect means of determining the distance to a star. The most obvious was to find a way of determining the actual intrinsic brightness of the star and compare it to the apparent brightness as we see it. If, for example, we knew that a certain star had the same actual brightness as our Sun it would be possible to calculate how much further away it was by comparing its apparent brightness to the Sun's. The problem of course is how to determine intrinsic brightness. This is the subject of the next part of our story. However, these indirect methods have to be 'calibrated' by testing them on stars whose distances are known through direct methods, and for this reason it is most important to use stellar parallax methods on as many stars as possible.

One limitation to parallax measurements is the distortion, or 'shimmer', of the star's image caused by the Earth's atmosphere. In 1989 the European Space Agency launched a satellite called Hipparcos to measure the stellar parallax of as many stars as possible from above the atmosphere. Hipparcos is an acronym for 'high-precision parallax-collecting satellite', but the name also honours the ancient Greek astronomer Hipparchus, who was one of the first to map the stars systematically. Hipparcos measured the parallax of more than 100 000 stars to an accuracy of 0.001 arcsec, in other words out to nearly 1000 pc. This may seem a vast distance, but given that our galaxy is about 50 000 pc wide and that we are about 8500 pc from its centre, you can see that a lot of stars are still 'out of range'.

Starlight—how bright?

Astronomers measure the apparent brightness of stars on a scale that actually originated with Hipparchus in the second century BCE! He called the brightest stars he could see 'first-magnitude' stars (+1), those about half as bright 'second-magnitude' (+2) and so on to those barely visible, which were 'sixth-magnitude'. When astronomers sailed into the southern hemisphere, they discovered brighter stars and so the scale had

Physics file

What is a parsec (pc)? Astronomers define the **parallax angle** as the angle subtended by the radius of the Earth's orbit (1 AU) as it would be seen at the distance of the star—as shown in Figure 6.37. The total circumference of a circle of radius r would be $2\pi\rho$.

As 1 arcsec is $\frac{1}{3600}$ of a degree, and a degree $\frac{1}{360}$ of a full circle, we can see that 1 AU must be:

$$\frac{1}{3600 \times 360} = \frac{1}{1\,296\,000}$$

of the full circumference; that is:

$$1\text{ AU} = \frac{2\pi r}{360 \times 3600}$$

You can show that this leads to r = 206 265 AU, which is indeed the conversion factor from pc to AU.

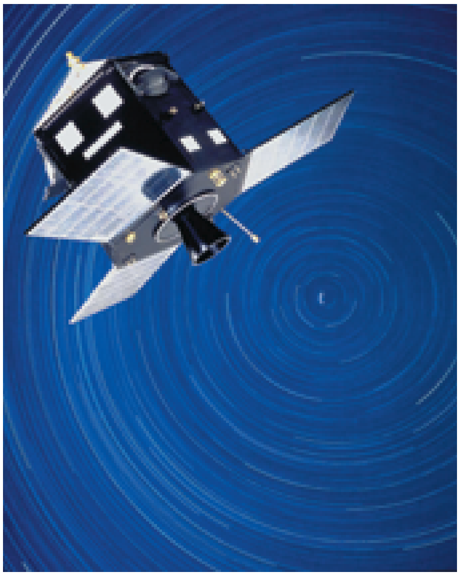


Figure 6.37
The telescope on the Hipparcos satellite was able to measure the parallax of stars to one-thousandth of an arcsecond.

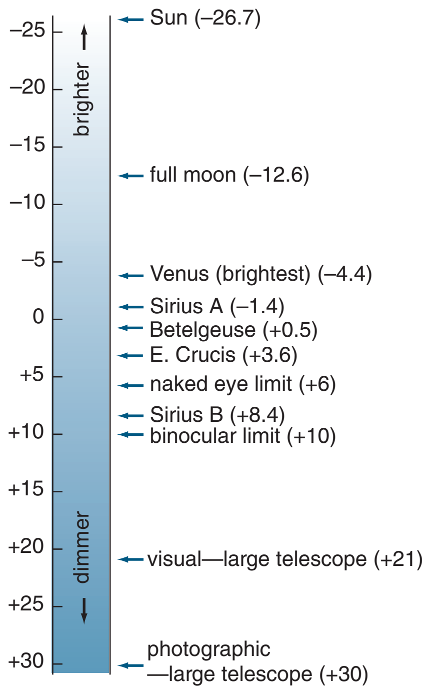


Figure 6.38

The apparent magnitude scale is a concoction of an ancient Greek scale with a 19th century mathematical redefinition! Visible stars range from –1.44 (Sirius A) 'down' to about +6 for stars barely visible under the best conditions.

Figure 6.39

The brightest star in our sky—our Sun. It is a huge ball of very hot, very dense gas.

to be extended 'upwards' to 0 magnitude and then –1 magnitude and so on. When the telescope was invented, the scale went 'downwards' to +7 and beyond. The scale is referred to as the *apparent magnitude* scale. Don't be confused by the fact that the scale seems 'backwards'—dimmer stars have a numerically higher magnitude.

In the 19th century when astronomers were better able to measure the brightness of stars, they defined apparent **magnitude** more precisely by saying that a difference in magnitude of 5 corresponds exactly to a factor of 100 times in apparent brightness—which agreed roughly with the old values. In other words, it would take 100 stars of magnitude +6 to equal the brightness of one star of magnitude +1. Mathematically this means that each level of magnitude represents a change in brightness of about 2.5 times (instead of Hipparchus's 'double'). In other words, it would take 2.5 stars of magnitude +6 to equal the brightness of a single +5 star, for example. Or it would take $2.5 \times 2.5 = 6.3$ stars of magnitude +6 to equal the brightness of a +4 star—and so on. Figure 6.38 shows the full extent of the scale with some examples. You might like to check with your calculator that it would take about 13 billion Sirius stars to equal the apparent brightness of the Sun!

Our favourite star

The brightest and most important star in our sky is the Sun. Without it, the Earth would be a frozen wasteland of ice and rock. Our planet is at just the right distance to receive enough radiation to keep water liquid. This in turn enables the sea and atmosphere to distribute heat around the Earth in such a way that the planet is habitable. Any closer and we would have the nightmarish greenhouse of Venus; any further away and the oceans would freeze. For humans, the Sun has always been an object of fascination, even worship. We will look at it through the eyes of the astrophysicists, who specialise in trying to understand the processes that drive the Universe.

Our best-known star

It was only after Galileo and Newton that we began to realise that the stars were actually Sun-like objects a very long way away. The best way to learn about stars, then, was to look at the closest one. After Galileo's discovery of sunspots it was realised that the Sun rotates on its axis. Curiously enough, however, the equator of the Sun was seen to have a period of about 25 days while regions of higher latitude took several more days for a full cycle, indicating that the Sun is not a solid body like the Earth. Because it is obviously so hot, it was assumed to be gaseous.

While the relative distances between the Sun and planets were known quite accurately as a result of the work of Tycho Brahe and Johannes Kepler, it was a far harder task to find the actual distances. The problem was to find the *scale* of the 'map' of the solar system. The only way to do that was to use a triangulation method somewhat akin to the parallax methods we have discussed for the stars. If a planet can be observed from opposite sides of the Earth at the same time, against the background of stars, the distance to the planet can be calculated in terms of the Earth's known diameter. As this distance will also be known in astronomical units, the scale can be found. This was attempted with some success in the late 1800s, but it was not until the advent of astronomical photography and better transport that reliable determinations could be made. One of the first sightings used the asteroid Eros, which came within 0.27 AU of the

Earth in 1901. The advantage of using Eros was that because it was small it was easier to pinpoint its location at a certain time. In the later 20th century it became possible to bounce radar pulses off the closer planets and hence find the time for light to travel the distance to the planet. This enabled very accurate determinations of the scale of the solar system map, and hence the value of 1 AU.

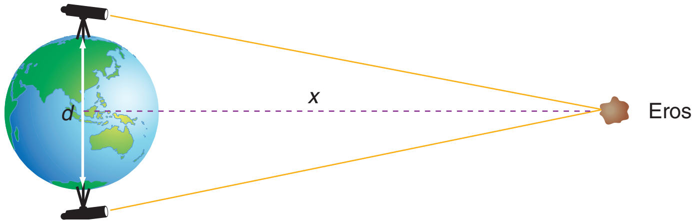


Figure 6.40

In order to determine the actual, as distinct from relative, distances in the solar system, some must be measured in terms of a known distance—such as the Earth's diameter.

Knowing the distance to the Sun, it is possible to determine its size very accurately. Furthermore, by using Newton's law of gravitation, it is possible to determine its mass. These two values can be used to determine the average density of the Sun. These values are summarised in Table 6.4. We can see that the Sun is gigantic compared with the Earth, except for its average density, which is only about 1.5 times that of water. On the other hand, as the Sun is basically composed of gas, we can see that it is a very dense gas!

Practical activity

42 Solar constant

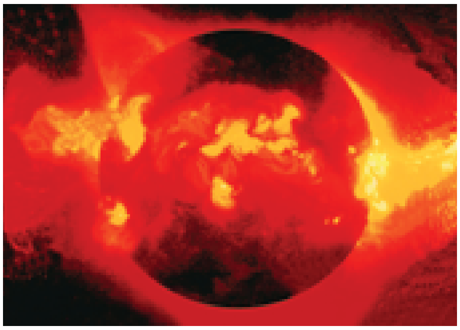


Figure 6.41

The Sun, photographed by a space-based X-ray telescope. The bright patches are regions that are producing solar flares.

table 6.4 The physical properties of the Sun

Property	Value
Average distance to the Sun (1 AU)	1.496×10^{11} m (390 times as far as the Moon)
Angular diameter from Earth	0.5° (same as Moon)
Mass of Sun	1.99×10^{30} kg (300 000 times the Earth)
Diameter of the Sun	1.4×10^{9} m (10^9 Earth diameters)
Average density of the Sun	1.4×10^{3} kg m^{-3} (Earth 5.5×10^{3} kg m^{-3})

From where does the energy come?

For a long time this was one of the great mysteries of physics. It was known that life had existed on the Earth for at least several hundred million years and so that meant that the Sun must have been radiating energy at much the same rate for at least that time. However, any known mechanism for producing heat could not possibly have generated so much energy. If the Sun was fuelled by even the most efficient chemical fires, it would have lasted only around 10 000 years. The English physicist Lord Kelvin (of absolute temperature fame) and the German physicist Hermann von Helmholtz suggested that vast amounts of heat would be generated from the enormous weight of gas collapsing into the Sun. As the gas falls, the potential energy is converted into kinetic energy—just as compressing air in a bicycle pump gives it kinetic energy, which heats it. This was good physics, but again it could not possibly last for the billions of years that the Sun has been producing energy. However, this process, now known as Kelvin–Helmholtz contraction, does turn out to be important in the formation of new stars.

Physics file

table 6.5 Astrophysical constants

Constant	Symbol	Value
Mass of Earth	M_E	5.97×10^{24} kg
Radius of Earth	R_E	6.38×10^{6} m
Mass of Sun	$M_\odot$	1.99×10^{30} kg
Sun–Earth distance	1 AU	1.496×10^{11} m
Luminosity of the Sun	$L_\odot$	3.86×10^{26} W

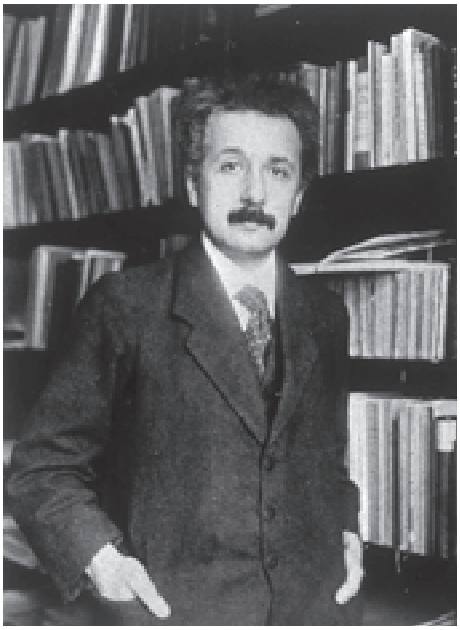


Figure 6.42

Einstein discovered that mass and energy were interrelated by his famous equation $E = mc^2$.

Physics file

Why is so much energy released when hydrogen atoms fuse to form helium? We can think of it this way: if two ordinary magnets are placed near each other so that they attract, and then let go, they will 'fall' together, gaining speed. When they collide, this kinetic energy is released as a little heat and sound energy. Something like this occurs when hydrogen nuclei 'fall' together to form a helium nucleus. The nuclear force that pulls the protons and neutrons together to form helium is enormously strong, but operates only over a short range (otherwise all hydrogen atoms would collapse together, forming helium). In order for the nuclear force to take effect, the huge repulsion between the positively charged protons has to be overcome. This is where temperature is important. Only with the sort of kinetic energy the particles have inside the core of the Sun can the particles come close enough, against the electrical repulsion, for the nuclear force to take over. When they do, however, they 'fall' together with incredible force, releasing huge amounts of energy.

Reasonably straightforward calculations based on the mass and energy output of the Sun showed astrophysicists that the amount of energy being produced for each atom in the Sun was hundreds of millions of times greater than the energy produced by each atom in chemical reactions. Something very different was going on in the Sun! The clue came with Einstein's theory of special relativity in the years after 1905: we all know that he came up with the equation $E = mc^2$. A reasonable interpretation for our present purposes is that there is a huge amount of energy locked up in mass. The factor c^2 is the speed of light squared and so is a very large number (9.00×10^{16} m^2 s^{-2}). This means that if this energy can be released, a huge amount of energy will be produced at the expense of a small amount of mass.

This suggested to the astrophysicists that the Sun may somehow be tapping this so-called 'mass energy'. But what was the process involved? Because of the enormous amount of energy radiating from the Sun, and assuming the energy was being produced in the centre, it could be calculated that the temperature at the centre of the Sun must be millions of degrees. This would mean that the atoms at the centre would be stripped of their electrons and there would be just a frenzied mass of nuclei and electrons flying around at enormous speeds. Because the density of the Sun was relatively low, it was assumed that it must be composed mostly of hydrogen with perhaps some other light atoms. It was already known that the Sun contained helium; in fact it had been discovered in the Sun before it was known on Earth. In the 1920s British astrophysicist Robert Atkinson suggested that under the conditions at the centre of the Sun the hydrogen nuclei might 'fuse' together, creating helium nuclei. If this were the case, huge amounts of energy would be released.

That the energy released in this 'hydrogen fusion' is so much more than that released in chemical reactions can be seen from the fact that chemical reactions involve forces between the electrons around the outside of the atom, whereas the nuclear fusion involves the electrical (and nuclear) forces between the particles in the nucleus. As the electrical force increases with the inverse square of the distance between charges (that inverse square law again!) and the charges in the nucleus are about 10 000 times closer than those in the outer regions of the atom, the forces between nuclear particles are about $10\,000^2$ or 100 million times larger—and so is the energy involved. This is the source of the energy that powers the Sun.

Figure 6.43

Nuclear reactions involve forces that are about 100 million times as great as those involved in chemical reactions. This is about the same ratio as the force required to keep a jumbo jet in the air to that required for a mosquito to fly.

✓ Worked Example 6.3A

The Sun is giving out about 4×10^{26} J of energy every second as visible and invisible radiation. At what rate is it losing mass due to this energy loss? How significant is this mass in proportion to the total mass of the Sun?

Solution

This energy comes from the fusion of hydrogen into helium with a corresponding loss in the potential energy of the nuclei. This loss of energy will correspond to a mass loss of:

$E = 4 \times 10^{26}$ J $\qquad E = \Delta mc^2$

$c = 3.00 \times 10^8$ m s^{-1} $\qquad \Delta m = \dfrac{E}{c^2} = \dfrac{4 \times 10^{26}}{(3.00 \times 10^8)^2}$

$$= 4 \times 10^9 \text{ kg s}^{-1}$$

or about 4 million tonnes every second. As the mass of the Sun is 2×10^{30} kg this is:

$\Delta m = 4 \times 10^9$ kg $\qquad \dfrac{\Delta m}{m_\Theta} = \dfrac{4 \times 10^9}{2 \times 10^{30}}$

$m_\Theta = 2 \times 10^{30}$ kg $\qquad = 2 \times 10^{-21}$ of the Sun's mass per second

At this rate it would take:

$\Delta m = 4 \times 10^9$ kg s^{-1} $\qquad \Delta t = \dfrac{m_\Theta}{\Delta m} = \dfrac{2 \times 10^{30}}{4 \times 10^9} = 5 \times 10^{20}$ s

$m_\Theta = 2 \times 10^{30}$ kg $\qquad = \dfrac{5 \times 10^{20}}{365.25 \times 24 \times 60 \times 60} = 1.6 \times 10^{13}$ years

or about 16 000 billion years to consume itself.
Actually only a little under 1% of the mass of hydrogen nuclei is lost in fusion so that reduces this value to under 100 billion years. However, as the Sun is 'only' about 5 billion years old we are in no danger of it running out of fuel for some time!

Modelling the Sun

It was the combination of Einstein's theory of relativity and the improved model of the atom (Bohr's quantum atom was suggested in 1913) that enabled astrophysicists to understand the processes occurring in the Sun. Today, computer models have enabled us to gain considerable insights into the mechanisms that keep the Sun 'burning' in our sky. These models are based on the same laws of physics that govern the transfer of heat from a log fire to you, or the laws that tell us that the pressure will increase as we dive into the ocean, and so on. The basic principle used in modelling the Sun is that any part of it must be in what is called *hydrostatic equilibrium*. That means that on any 'piece' of Sun, the inward pressure from the weight of all the material above it must be balanced by the outward pressure of the radiation released by the nuclear reactions in the core below.

The origin of the energy is rather different from the log fire, and the Sun's interior is much hotter than the ocean, but the same basic laws of

Physics file

Is mass converted into energy? It is often stated that in nuclear reactions 'mass is converted into energy'. This is really an oversimplification of Einstein's ideas. There is no mass 'shaved off' the particles and somehow mysteriously turned into energy. What is true is that the total mass of the particles in a helium nucleus is a little less than the total mass of the equivalent particles in hydrogen nuclei. However, they are exactly the same particles as before, except that they now have a little less total potential energy than previously (that was what went off to power the Sun). Einstein showed that mass is actually a property, not of the individual particles, but of the *system of particles* including the energy bound up in the forces between them. As some of the energy has been lost, so also has the equivalent mass (as given by Einstein's equation).
In fact there is no fundamental difference between a nuclear reaction and a chemical reaction in this sense. A group of hydrogen and oxygen atoms will also have a slightly greater mass than the equivalent group of water molecules formed when the hydrogen 'burns' with the oxygen, releasing energy. It is just that in this case the mass equivalent of the energy released is so tiny it is completely unmeasurable. But $E = mc^2$ applies just as surely in this case as in the nuclear case.

Practical activities

32 Sunlight's intensity and reflectivity of the Earth's surface
46 The sun in the day sky

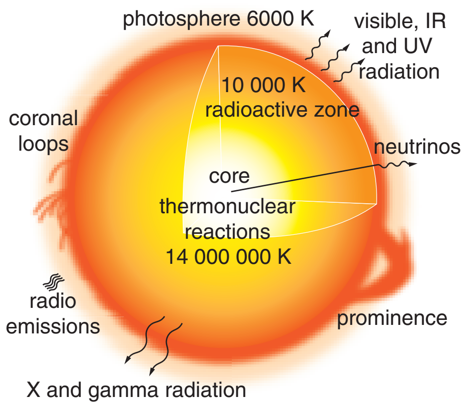


Figure 6.44

The astrophysicists' model of the Sun. The nuclear fusion occurs in the central core, from which energy slowly travels by radiative diffusion out to about $0.7R_\odot$ where the nuclei and electrons recombine. From here out the energy transfer is by convection.

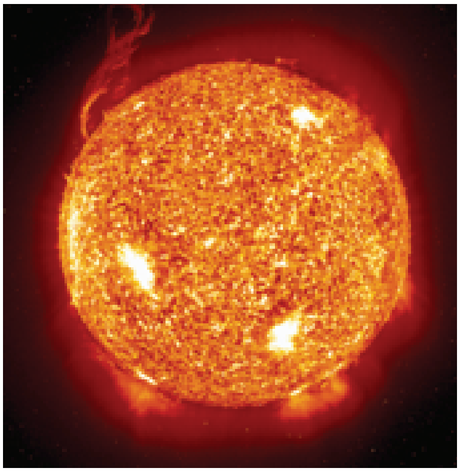


Figure 6.45

The granular appearance of the Sun's surface is due to the convection currents reaching the surface. Each granule is about 1000 km wide. They last a few minutes before disappearing and re-forming.

physics apply. The models take an imaginary piece of the Sun, and put in assumptions about quantities such as its mass, temperature, heat content and ability to transfer heat, the pressure gradient across it and much more. Then the conditions to ensure that it is in hydrostatic equilibrium are determined. This is repeated over and over until the whole Sun is accounted for, each 'piece' interacting with its neighbouring pieces. A computer then 'runs' the Sun to see what would happen. If the computer Sun blows up or dies out, we know some of our assumptions were wrong! Many such scenarios can be tested until they are found to be consistent with the known facts about the Sun, such as its mass, size, energy output and surface temperature—and the fact that the Sun has produced a very constant output over the last few billion years.

These models suggest a number of interesting 'facts' about the Sun. As it is impossible to test them, they must be seen as theoretical predictions rather than facts, but astrophysicists have a lot of confidence in their model. The main features of the model are that the nuclear fusion occurs in a zone that extends from the centre out to about 0.25 of the radius ($R_\odot$). The temperature in this region is above 10 million degrees, the density about 160 000 kg m^{-3} (that is 14 times the density of lead!), and the pressure about 340 billion times the Earth's atmospheric pressure. The energy flows outward from this zone by a combination of convection and radiative diffusion. The convection is just like that from a room 'convection heater'—hot air rises, cooler air sinks. Radiative diffusion is a process in which light bounces around, transferring heat in the process. Radiative diffusion is the main mechanism for energy transfer out to about $0.7R_\odot$, where convection takes over. At this point the temperature is down to about 1 million degrees and the protons and electrons come together to form hydrogen atoms, which absorb the light more effectively, and so the radiative diffusion becomes less effective. The density is only 80 kg m^{-3} (much less than water, but about 60 times that of the air around us) and the pressure is only about 10 million 'atmospheres'! Most (99%) of the Sun's mass is below this level.

By the time we reach the surface (in this very imaginative journey) the temperature is down to the 5800 K we 'see' as we look from our more comfortable position on Earth. There is of course no real 'surface', just a point where the churning hot gases start to sink again as they lose their energy by radiation out into space. The motley appearance of the Sun in detailed photographs (such as Figure 6.45) is due to the convection currents reaching the surface, cooling and then sinking again. This may all sound like a rapid process but in fact it has been calculated that energy produced by the nuclear fusion in the core takes about 170 000 years to travel out to the surface. Then it only takes 8 minutes to reach the Earth. The light energy we see by today was generated inside the Sun when humans were just beginning to diverge from the apes!

The Sun's atmosphere . . .

As the Sun is all gaseous, there is no real surface, but the thinner outer layers are often referred to as its atmosphere. The layer that we see in photographs such as Figure 6.45 is called the *photosphere* because it is the hot (5800 K), thin layer from which the Sun's visible light is emitted. In fact it is only about 400 km thick, in the sense that we can see only about that far into it. Note that we are doing very well to see that far in the Earth's atmosphere! Compared with the Sun's diameter, however, the photosphere is very thin and so we see it as a sharp, clear surface.

There are two more layers above the photosphere. Normally we can't see them, but during an eclipse when the Moon blocks the light from the photosphere, we can see a pinkish layer just above the photosphere. This is called (because it is coloured) the *chromosphere*. It is extremely thin, only about one hundred-millionth as dense as our atmosphere, and extends to about 2000 km above the photosphere. The chromosphere glows with coloured light for reasons we shall discuss shortly. Curiously enough, the temperature of the chromosphere increases with height, reaching about 25 000 K at the top.

The top layer of the Sun's atmosphere is the *corona*, which is the spectacular, but faint, whitish glow around the Sun that becomes visible only in an eclipse (or from a specially designed telescope in space). It is so thin that we would regard it as a very high vacuum if we produced it in the laboratory. However, it extends for millions of kilometres out from the Sun's surface. It consists of highly ionised ions (i.e. atoms that have lost many of their electrons) and free electrons at temperatures over 1 million degrees. Because particles at this temperature are moving so fast, some of them escape the Sun's gravity and stream out into space. This is the origin of the 'solar wind' which, when it interacts with the Earth's magnetic field and atmosphere, creates the auroras seen in polar regions. While the amount of material streaming from the Sun in the solar wind amounts to about a million tonnes every second, at this rate the Sun will only lose well under 1% of its mass in its entire lifetime. Most of the solar wind consists of protons (hydrogen nuclei) and electrons, but a small percentage contains ions of heavier elements.

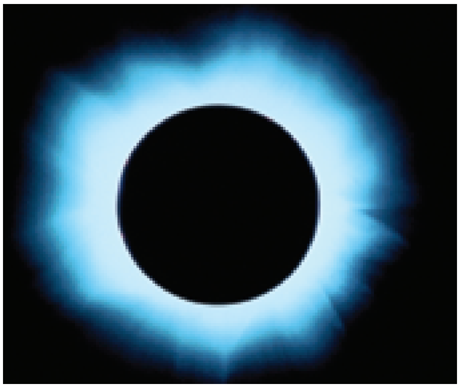


Figure 6.46

The corona we see around the Sun in an eclipse is there all the time but masked by the brilliance of the Sun. It is the origin of the 'solar wind' that, among other things, causes the beautiful auroral lights on Earth.

Practical activities

43 Watching the night sky
47 The phases of the Moon
49 Night sky exercises in astronomy

. . . and its magnetic field

The Sun's atmosphere and magnetic field interact to produce a number of spectacular effects including sunspots, solar flares and coronal mass ejections. The Sun's magnetic field appears to be much more dynamic than the Earth's. While the Earth's field may reverse a few times in a million years, the Sun's does it on a regular 11-year cycle. This is the underlying reason for the well-known 11-year sunspot cycle. Models of the Sun's magnetic field suggest that it becomes distorted because of the differential rotation of the Sun, in effect producing 'knots' that cause large disturbances in the photosphere, which can last several days. These disturbances are cooler, at 4300 K, and look dark against the average 5800 K surface. We see them as sunspots. Eventually the magnetic field gets itself in such a knot that it reverses and repeats the cycle all over again.

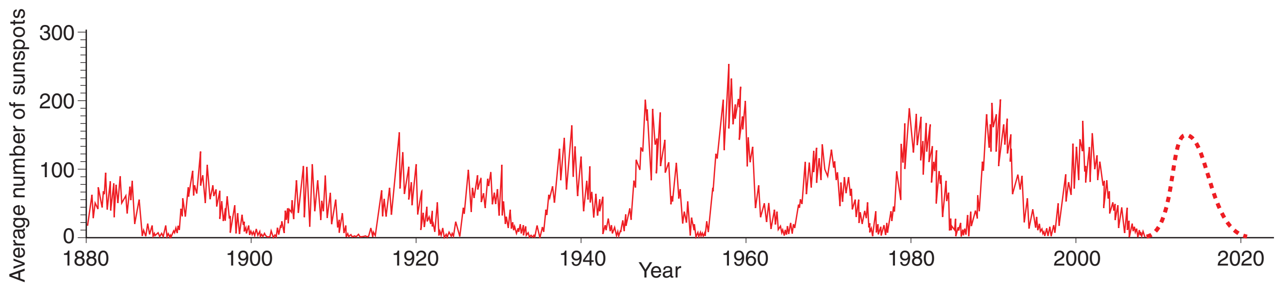


Figure 6.47

Sunspots come and go with an 11-year cycle, with another maximum predicted for 2018.

The 'magnetic storms' that give rise to sunspots are also the origin of the solar flares and other emissions from the Sun's surface that affect the chromosphere and corona. A *coronal mass ejection* can blast a billion tonnes of coronal gas into space at speeds of hundreds of kilometres per second. These events can produce large increases in the solar wind which, when they reach the Earth a few days later, are hazardous to astronauts and even to communication and electric power systems on Earth.

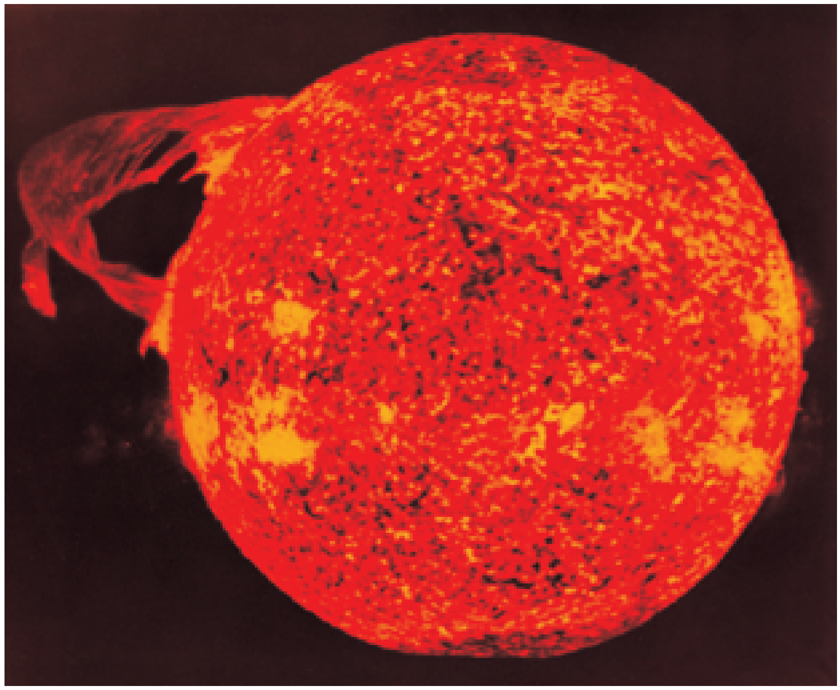


Figure 6.48

Magnetic storms on the Sun's surface create solar flares that result in 'coronal mass ejections'. These storms result in increased solar wind activity that can be hazardous for astronauts and can sometimes cause interference to electrical systems on Earth.

6.3 SUMMARY To the stars

- The distance to stars can be measured by the parallax movement they show as a result of the Earth's revolution around the Sun, but even the largest parallax found is less than 1 arcsec.
- 1 pc is the distance to a star that would show 1 arcsec of parallax. 1 pc is about 3.26 l.y. or 206 265 AU.
- The distances to more than 100 000 stars have been measured by stellar parallax, but most stars are too distant for this method.
- The relative sizes of the orbits in the solar system have been well known since Kepler, but it required triangulation using the Earth's diameter to determine the actual distances.
- The distance to the Sun, its size, mass and density have been determined very accurately.
- The source of the Sun's energy was a mystery until the nature of nuclear forces was understood.
- The Sun produces energy by the fusion of hydrogen nuclei to form helium nuclei.
- The light we see from the Sun is emitted from the photosphere, the temperature of which is 5800 K.
- The solar wind originates in the Sun's uppermost atmosphere, the corona, which we can see only during an eclipse. The solar wind influences the Earth's upper atmosphere.

6.3 Questions

1 An arc second is a very small fraction of a degree. How many arc seconds are there in 1°?

2 The angular diameter of the Moon is about 0.5°. How many arc seconds is that?

3 How far away would a 5-cent coin need to be to subtend an angle of 1 arcsec?

4 Explain why the scale for the apparent magnitude of the brightness of a star has negative values for the brightest stars.

5 Even before the process of nuclear fusion was known, astrophysicists deduced that the temperature inside the Sun must be millions of degrees although the surface temperature was 'only' about 5800 K. Can you suggest reasons why they would have come to this conclusion?

6 Why is it that the nuclear reactions involving hydrogen in the Sun can produce so much more energy than hydrogen burning in oxygen on the Earth?

7 Astrophysicists model the Sun by using powerful computers. Briefly outline the basic assumptions that are programmed into the computers.

8 Briefly describe the processes that bring the heat generated in the core of the Sun to the Earth.

9 Which of these statements best describes the nature of sunspots?
A Sunspots are due to disturbances in the nuclear reactions in the core of the Sun.
B Sunspots result from convection currents being formed in the Sun's outer layers.
C Sunspots result from magnetic disturbances in the photosphere of the Sun.
D Sunspots result from holes in the Sun's upper atmosphere (the chromosphere and corona).

10 Which layer of the Sun's atmosphere gives rise to the sunlight we see arriving on Earth?
A Photosphere
B Chromosphere
C Corona
D All layers

6.4 Fundamentals of astronomy

There is probably no more spectacular natural display than the heavens as seen from a dark setting away from city lights. Imagine yourself lying awake (in a warm sleeping bag) on a mountain top on a clear moonless night. The Milky Way, our galaxy, is particularly beautiful. In the southern hemisphere we are privileged to be able to see into its heart. It may look like a milky band of cloud, but if we look with binoculars we see that it is actually composed of myriad stars and nebulae (seen as small fuzzy patches). In late summer, early evening, the two brightest stars in the night sky—Sirius and Canopus—are high overhead: Sirius a little north of the zenith (directly overhead) and Canopus a little to the south. Just to the north-east of Sirius we find the constellation of Orion the Hunter with his distinctive belt and sword—well, distinctive when seen from the northern hemisphere. From our part of the world Orion appears upside down and we often refer to this group of stars as the 'Saucepan'.

Practical activities

43 Watching the night sky
47 The phases of the Moon
49 Night sky exercises in astronomy

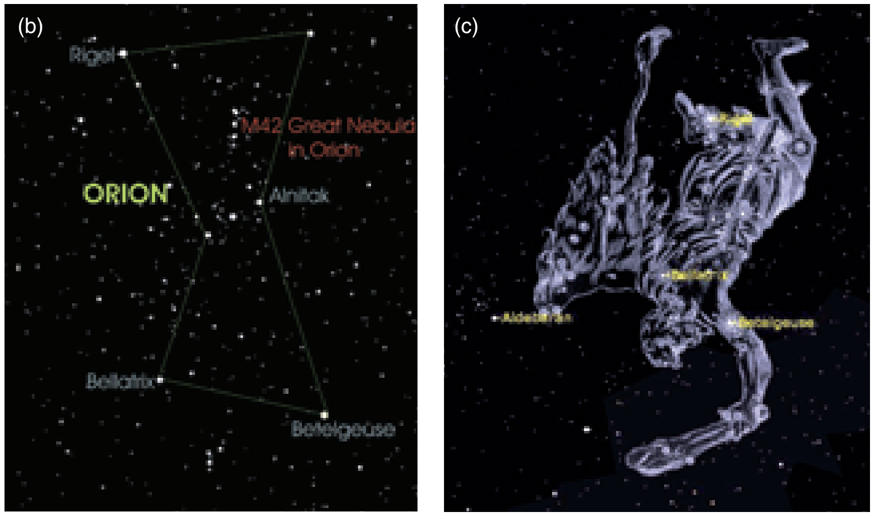


Figure 6.49
(a) Orion as seen from the southern hemisphere. Note the colours of the stars. (b) The Great Nebula in Orion can be seen with binoculars. (c) Orion the Hunter is upside down when viewed from the southern hemisphere.

Look a little more closely at the two bright stars that make up Orion's shoulders (Betelgeuse and Bellatrix). They are the two stars a little below the Saucepan. You will notice that Bellatrix is distinctly bluish while Betelgeuse is quite red. Compare them with Sirius (a little to the south and east of Orion), which is quite white. The colours of the stars are of great importance to astronomers. We can only know the stars by the light we receive, and the colour of this light is an important clue to their nature. In the past century many techniques have been developed to analyse this light and to extend our vision into the invisible 'light' well beyond the visible spectrum.

You may also be able to see some planets. Planets tend to be brighter than most stars and do not 'twinkle' (shimmer due to atmospheric effects). Venus, the so-called 'evening star' (or 'morning star') can be particularly stunning. It is not really a star of course, but it is the brightest object in the night sky, apart from the Moon and the International Space Station. Mars, Jupiter and Saturn can be quite spectacular also and are often brighter than most stars. Through binoculars we can make out Jupiter's moons, and, with a small telescope, Saturn's rings. As we watch the stars from our mountain top we notice that they are gradually moving, rising from the east and setting in the west, as do the Sun and the Moon. In fact, when we wake at 4 a.m. the next morning the starscape is quite different. Orion has set and Scorpius is now high above us. The Southern Cross, which in the evening was low in the south-east, is now quite high in the southern sky.

Figure 6.50
The Southern Cross (right) and the Pointers (left) as seen when it is high in the sky

It is sometimes a surprise to those of us who live in cities and rarely notice the night sky to discover what the ancients knew as common knowledge. If we were to stay awake all night and *carefully* watch the sky, we would find that the stars rotate a full circle each day about a point or 'pole' in the sky. This point is the south celestial pole (SCP) in the southern hemisphere and the north celestial pole (NCP) in the northern hemisphere. From Australia we can see the SCP (see Figure 6.51) but not the NCP. In fact, as we shall see, the altitude of the SCP above our southern horizon is equal to our latitude.

Figure 6.51

The diurnal (daily) motion of the stars. In this 10-hour exposure from the Anglo-Australian Observatory at Siding Springs the stars have rotated 150° around the south celestial pole.

Practical activities

44 Measuring the night sky: alt-azimuth
45 Measuring the night sky: equatorial coordinates

Physics in action — A strong recommendation

Although it would be possible to complete this study entirely from within the classroom, we strongly suggest that at some point you try to take a trip to a country location where you will have a clear view of the night sky away from the light pollution of cities and towns. The ideal would be a trip to the wheat-belt, but even a 50 km trip away from the city lights can provide a much better view of the night sky than is possible from the outer suburbs. Moonlight will also overwhelm many of the stars, so choose a night when the moon will not be up in the late evening. However, the crescent moon, just after new moon, can be a wonderful sight in the early evening. Take along a star chart of some sort (and a dim red torch to read it by as red will not spoil your night vision) as well as the 'planet rise and set' table published in the daily paper. The latter will enable you to identify any planets, as they cannot be shown on a star chart. Even better, take an astronomy yearbook which lists all sorts of information and suggestions for viewing for the current year (see Table 6.6).

All you really need are your own two eyes, but a pair of binoculars can significantly enhance the experience. The lens diameter of the binoculars is more important than the magnification. A pair of 7×50 (magnification × diameter) binoculars will give a brighter image than an 8×30 pair, for example. It is well worth keeping them steady on a tripod or other firm support. Short black cardboard tubes on the front of the lenses will protect them from stray light, as well as from dew on cold nights. Make sure the lenses are clean as the slightest smear from a finger on the lens will tend to put halos around the stars.

If you are appropriately equipped, it can be a wonderful experience to sleep outside 'under the stars'. Try to wake up several times through the night to see the changes that have occurred during the night. The early morning sky will be quite different from the evening sky.

Table 6.6 lists some of the objects particularly worth looking for, but don't neglect to simply enjoy the experience as a whole. Then spend some time identifying the brightest stars and the more obvious constellations. Notice the different colours of the stars (binoculars are helpful here), and the planets if they are visible. The Milky Way is our 'home galaxy', and on a clear, dark night can be a wonderful sight indeed. Because we are looking at it from within, it stretches right around our sky. We are lucky that the centre of our galaxy, the brightest part of the Milky Way, is actually in the southern sky.

table 6.6 Some astronomical sights

With the naked eye	• Use a star chart to identify the major constellations. Start with the Southern Cross and Pointers (part of Centaurus), Orion (with the 'Saucepan') and Scorpius. Then look for some key stars such as Sirius, Canopus, Aldebaran, Achenar and Formalhaut. • The Milky Way—our galaxy: It is particularly brilliant in winter and spring evenings (or the early morning in autumn). The centre of the galaxy is in the region near Sagittarius and Scorpius, so try to observe it when that area is high in the sky (i.e. in the evenings from July to September). • The Large and Small Magellanic Clouds: The LMC and SMC are patches of light near the south celestial pole. They are quite easily seen in a dark sky and look like pieces of the Milky Way that have 'broken off'. Indeed they are actually satellite galaxies of the Milky Way. • The colours of stars and planets: Look at the colours of the stars in Orion. Betelgeuse is one of the reddest stars while Bellatrix is one of the bluest. Because of the brilliance of stars it is not easy to see the colour with the naked eye; but if you have the chance to take a photograph, the colour will be more obvious. • The wandering planets: In one evening you will not see the planets moving among the stars, but if you can spread your observations over weeks or months you will see that the planets appear to move among the stars. If you are lucky you may notice that the direction of a planet's motion, with respect to the stars, changes and starts moving 'backwards' for a while. This is the 'retrograde motion' which made it so difficult for early astronomers to explain the motion of planets. • Conjunctions between stars, planets and the Moon: Astronomy yearbooks list various conjunctions of planets with each other, stars or the Moon when the objects involved appear to come close. These are fun to watch and, except for those involving the Moon, occur over a period of days or weeks.
With binoculars	• In good conditions, four of Jupiter's moons may be visible through binoculars. Look for the small 'stars' in a line nearby and watch them over the course of a few days. • The colours of stars and planets are more obvious through binoculars. Look particularly at the colours of stars such as Aldebaran, Betelgeuse and those of the Southern Cross. • The larger craters on the Moon will be visible, particularly where the shadows are long near the edge of the dark region (the 'terminator'), so look for them at phases other than full. Also look for the mare (seas) on the Moon. • There are many star clusters, some quite open such as Coma Berenicis or the Pleiades (the Seven Sisters), which is the most famous star cluster in the sky, and others small and dense, such as the Jewel Box near the Southern Cross. They are a wonderful sight. • Look for the Great Nebula in Orion—the middle star in the sword (or Saucepan handle). • One of Galileo's great discoveries was the fact that Venus showed phases that changed with its apparent size. The crescent phases may just be visible in good conditions.
With a telescope	• A small telescope may enhance your view of any of the observations suggested above, but many people find observation with binoculars just as satisfactory.
Sky-gazing tips	• Find the rise and set times of the planets, Moon and Sun in the daily paper. • Obtain a 'starfinder' or 'planisphere': a cardboard or plastic disk that rotates showing the stars that will be in the sky at any particular time. (These are available from museum shops, Australian Geographic stores, telescope stores and some bookshops.) • A year guide such as *Astronomy 20XX Australia* (where XX is the current year), by Dawes *et al.* (Quasar Publishing), is helpful. • A number of publications list sky objects to observe through binoculars or small telescopes. A good one is *Space Watching*, published by Australian Geographic, but there are a number of others.

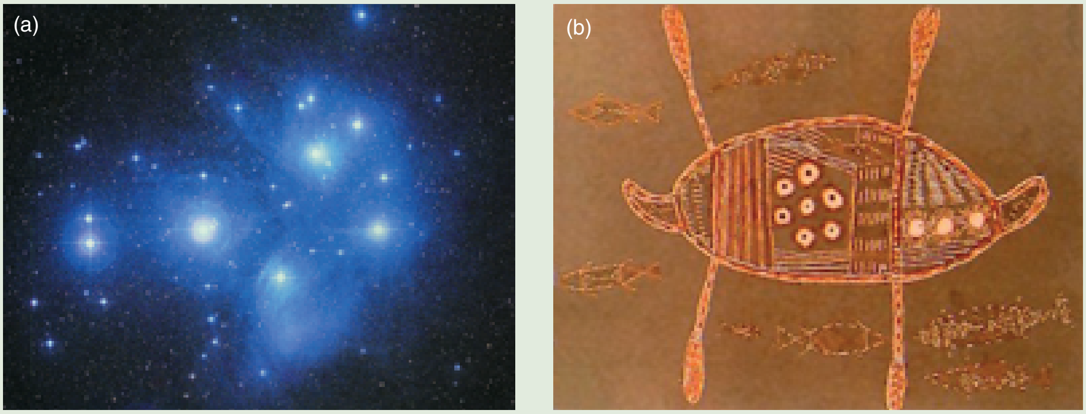


Figure 6.52

(a) The Pleiades (the Seven Sisters) is the best-known open star cluster and is a lovely sight through binoculars. (b) The Aborigines incorporated the Seven Sisters and Orion's belt into their Dreamtime legends. One story has Jarn the hunter chasing Marigu (which we know as the Seven Sisters) to try to catch a wife.

Early models of the Universe

Most of what we have said so far would make perfectly good sense to the ancient Greek astronomers, such as Pythagoras, who studied the heavens in great detail. Most of them accepted, however, that the Earth was a solid, immovable sphere at the centre of the great celestial sphere, which they believed to be truly heavenly, the realm of the gods. They argued on several grounds that the Earth was spherical. For example, it was well known that the sailor in the 'crow's nest' at the top of the mast could see further than those on the deck of the ship. By the same token, those on shore saw the hull of the ship disappear before the mast and sails. It was also known that as one travelled further north the Sun became lower in the south sky and the North Star rose higher. These facts were easily explained by assuming a spherical Earth. Perhaps their best proof was the curved shadow of the Earth on the Moon during a lunar eclipse.

It was very easy to imagine the stars rotating on their huge celestial sphere, which moved in a very stately and regular fashion. Inside the celestial sphere were more spheres for the other objects in the sky. The spheres containing the Sun and the Moon rotated regularly, but changed their orientation somewhat over the period of a year or a month, albeit in a very predictable fashion.

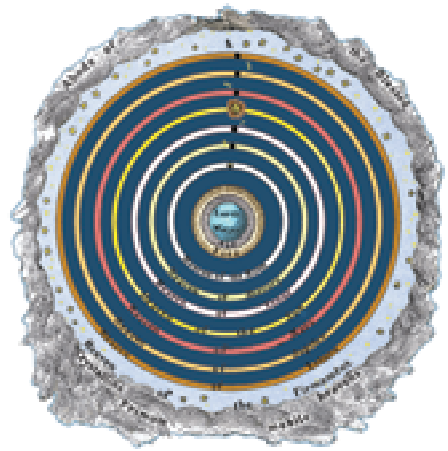


Figure 6.53

Aristotle taught that the Earth was at the centre of the Universe. It was surrounded by water, air and fire. Beyond that was the heavenly realm of the Moon, the Sun and the planets, which described perfect circles around the Earth. The celestial sphere, which surrounded all of this, was the region of the stars and the home of the gods.

Figure 6.54

As the Moon progresses into the Earth's shadow during a lunar eclipse, the shadow always appears curved. The Greeks realised that this meant that the Earth must be spherical.

Physics in action — Flight 143

In about 200 BCE, the Greek philosopher Eratosthenes found a way to estimate the Earth's size. He knew that at the summer solstice in the town of Syene (now under the present-day Aswan Dam in Egypt) the Sun was directly overhead—it cast no shadow of a vertical pole. At the same time in Alexandria, 5000 stades (a distance measure) to the north, a vertical pole did cast a shadow at midday. Measurements of this shadow and the height of the pole enabled Eratosthenes to calculate that the Sun was 7° away from being directly overhead. He therefore concluded that the distance between the towns represented 7/360 or close to 1/50 of the distance around the Earth's circumference. Thus the total circumference would be 50×5000 or 250 000 stades, which in today's units is roughly 42 000 km. Not a bad estimate when you compare it with today's value of close to 40 000 km!

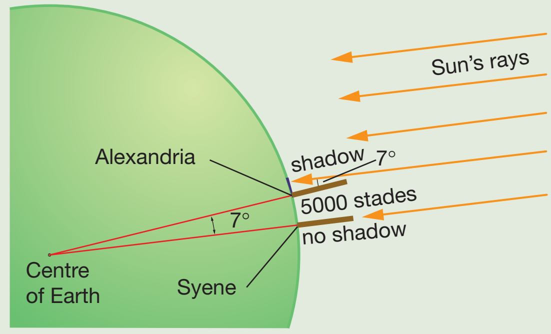
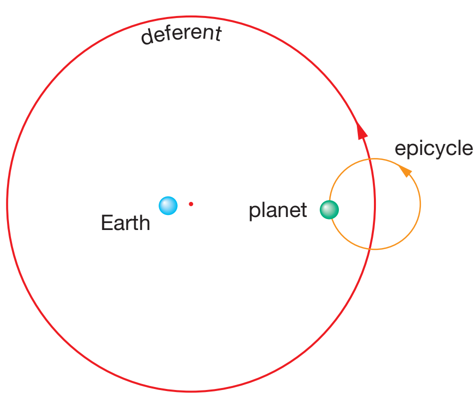


Figure 6.55
Eratosthenes estimated the Earth's circumference in 200 BCE by noting that while the midday Sun was overhead in Syene it cast a 7° shadow in Alexandria, 5000 stades to the north.

The planets were something of a problem, however. How could their rather irregular motion be reconciled with their heavenly and, therefore, 'perfect' nature? Circular motion, such as that of the stars, was seen as perfect and so it was thought that in some way the motion of the planets must be made up of some combination of different circular motions. In the second century BCE, Hipparchus suggested that the motion of the planets could be explained by a system of circles upon circles. Each planet was thought to move in a small circle called an *epicycle*, the centre of which moved in a larger circle called the *deferent*, which was centred on the Earth.

In the second century CE, Ptolemy, the last of the ancient astronomers, improved the accuracy of Hipparchus's system by developing a method for locating the centre of the deferent. He located the centre a little away from the Earth, and had the planet move around the circle at slightly varying speeds. Although a rather complicated system, it was based on the 'perfect circle' and was able to predict the future positions of the planets with quite a high degree of accuracy. In the third century BCE Aristarchus had suggested a model based on the Earth rotating around the Sun; however, the Ptolemaic system was seen as philosophically more satisfactory and thus lasted for over a millennium.



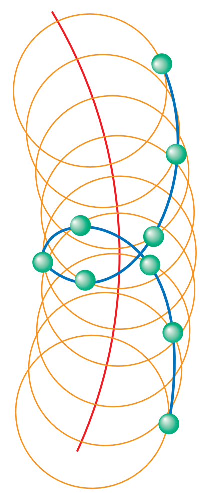


Figure 6.56
Ptolemy's system of planetary motion based on epicycles, the centre of which moved on a circle called the deferent centred near, but not on, the Earth. When on the inside of the deferent, the planet could exhibit retrograde motion as shown.

The Copernican revolution

In the Middle Ages, scientists, or 'natural philosophers', began to question the Ptolemaic system on the grounds that, although it worked well, it was not based on any underlying principle that could be used to work out the motion of any planet. The details had to be worked out to suit the observed motion of each planet separately. By this time, scientists had begun to realise that there seemed to be a principle in nature that favoured simpler explanations over complex ones.

Aristarchus had actually suggested a heliocentric (Sun-centred) model of the world back in ancient Greek times. He reasoned that it provided a simpler explanation for retrograde motion. However, to most people the idea of the Earth being in motion was an idea that seemed far harder to accept. In the early 16th century Nicolaus Copernicus, a gifted Polish mathematician, suggested that a Sun-centred model had some advantages over the Ptolemaic system. Not only did the heliocentric model explain retrograde motion more simply, it also explained why, for example, Mercury and Venus never appeared in the midnight sky while the other planets did.

The heliocentric explanation for retrograde motion can be seen in Figure 6.57. In the Copernican system, the further away the planet is from the Sun, the more slowly it moves around its orbit. As a result, the Earth 'overtakes' the outer planets at some stage in its orbit. As this happens, the outer planet will appear to travel backwards relative to the stars in the far background.

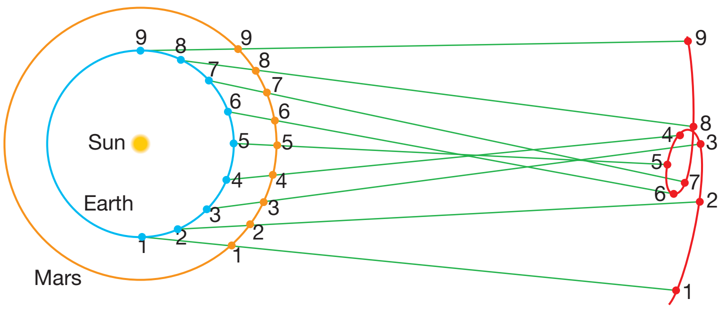


Figure 6.57
Using the Copernican explanation of retrograde motion, Mars appears to move backwards for a couple of months each year when the Earth 'overtakes' the more slowly moving planet.

It was well known that Mercury and Venus never moved more than a certain angular distance from the Sun. Venus is often referred to as the morning or evening 'star' because it is only seen within about 3 hours of sunrise or sunset. The Ptolemaic system had no explanation for this, but the Copernican heliocentric system had a very simple explanation: these two planets are closer to the Sun than is the Earth. They are said to be *inferior planets*. This was not a value judgement, simply a reference to the fact that the radius of their orbits was less than that of the Earth. All the other planets can, at some stage of their cycle, be seen high in the sky at midnight. This means that they must be further from the Sun than the Earth is, so they are *superior planets*. This is illustrated in Figure 6.58.

Copernicus was able to use trigonometry to calculate the radius of the orbits of the planets in terms of the radius of the Earth's orbit; that is, the distance to the Sun. This was simple for the inferior planets, as can be seen in Figure 6.58. At its maximum elongation from the Sun (46°) we can see that:

$$\sin 46° = \frac{R_V}{R_E}$$

$$R_V = R_E \times \sin 46°$$

Physics file

The 14th century English philosopher William of Occam first expressed the principle that nature seems to favour simple, elegant principles over complex ones. This became known as *Occam's razor* in reference to the fact that he suggested trying to shave unnecessary details from any explanation for a natural phenomenon. Ever since, scientists have found it a useful, although not infallible, principle. Albert Einstein actually based much of his thinking on Occam's razor, continually searching for simpler explanations for phenomena to do with light. Although the mathematics of relativity can be complex, its underlying principles are very simple.

Copernicus had no good way of determining the distance of the Earth from the Sun and so simply expressed the radii of the planet orbits in terms of **astronomical units**, or AU for short. 1 AU is the average distance of the Earth from the Sun. We now know it to be 150 million kilometres. A similar, but somewhat more involved, process enabled Copernicus to determine the distances to the superior planets. His values were quite close to the modern-day values, as shown in Table 6.7. Copernicus did not know about the three outer planets; their discovery had to await the invention of the telescope.

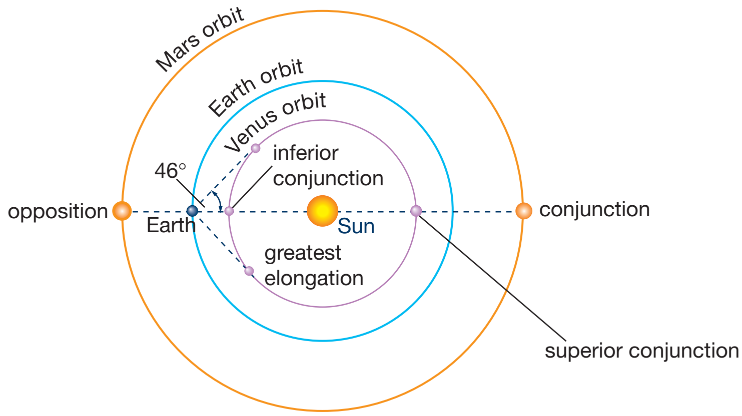


Figure 6.58

Inferior and superior orbits are illustrated by the orbits of Venus and Mars. Venus is never seen to be more than 46° from the Sun, but Mars can be seen in the midnight sky. When the inferior planets are in a line with the Sun and Earth they are said to be at inferior (same side as the Earth) or superior (behind the Sun) conjunctions. The equivalent points for the superior planets are called simply 'conjunction' when the planet is behind the Sun (and so not visible at night) and 'opposition' when the planet will be seen high in the sky at midnight.

While the Copernican system had a philosophical advantage in that it seemed conceptually simpler to some astronomers, it was no better at predicting astronomical phenomena than was Ptolemy's. In fact Copernicus initially found that his system gave worse results, and he had to resort to adding small epicycles to his orbits in order to achieve similar accuracy. His epicycles were not nearly as large as Ptolemy's, however, and were not needed to produce the retrograde motion. They simply had the effect of speeding up and slowing down the planet a little to make the orbits fit the known data.

table 6.7 Average distances of the planets from the Sun in AU

Planet	Copernicus's value	Modern value
Mercury	0.38	0.39
Venus	0.72	0.72
Earth	1.00	1.00
Mars	1.52	1.52
Jupiter	5.22	5.20
Saturn	9.07	9.54
Uranus	–	19.19
Neptune	–	30.06
Pluto*	–	39.53

* In 2006 the International Astronomical Union decided that Pluto should be downgraded from planet to dwarf planet.

Perhaps the most important aspect of Copernicus's achievement was that he had made a break from the past. He showed that despite the success of the Ptolemaic system it might not be the best way to understand the way the world works. In fact, Copernicus started a revolution. The publication in 1543 of his *De Revolutionibus Orbium Coelestium* ('On the revolutions of the celestial spheres') is now generally seen as just the start of the so-called 'Copernican Revolution' and is regarded as one of the great turning points, not just of science, but of history. It was a revolution that was to include some of the great names in physics, such as Galileo and Kepler. It also pitted the new science against the political authorities of the day, including the Church.

Practical activity

48 Computer simulation of the night sky

Observation and theory

Tycho Brahe was born in 1546, just a few years after Copernicus's death in 1543. He grew up in a wealthy Danish family and became a nobleman. His wealth allowed him to pursue his passion for astronomy. In particular, he wanted to catalogue the stars and the motion of the planets. In 1572 he experienced a truly amazing sight: a new bright star, even brighter than Venus, suddenly appeared in the constellation of Cassiopeia. Today we recognise that event as a supernova, the explosion and death of a massive star. At that time, however, the celestial realm was seen as permanent and unchangeable, the home of the gods, and so most people argued that the new 'star' must have been something much closer to Earth.

Brahe was determined to investigate. He figured that if the new object really was closer than the 'real' stars, then it should be possible to see it move against the background stars, as do the planets. This is known as *parallax* and is a familiar phenomenon in everyday experience. If you look out the window at something in the distance and then move your head from side to side, the distant object appears to move with you, relative to the closer objects, such as the window frame, which moves against you. You may have seen a rising Moon skipping along the treetops as you drive along a country road at night. The Moon is not moving at all (well, not much); it is your motion that makes the trees appear to move backward relative to the Moon (or anything that is further away than the trees).

With this in mind, Brahe observed the new star very carefully during the course of a single night and during the course of the next months as it faded from view again. He found no parallax between the star and the celestial sphere and therefore concluded that the star was indeed very far from the Earth; perhaps in the 'celestial realm' itself. Cracks were beginning to appear in the 2000-year-old view of a 'permanent and unchangeable realm of the gods'.

As a result of his careful work on the new star of 1572, Brahe was awarded a grant by the King of Denmark to build two state-of-the-art observatories. There were no telescopes yet, but there were very carefully built instruments with which to measure the positions of the stars and planets. Brahe decided that with careful observation he could test the Copernican theory of a moving Earth. If the Earth was indeed circling the Sun, surely we should notice some parallax movement in the positions of the stars, he reasoned. There was plenty of parallax movement between the planets and the stars, but this was the reason for believing that the planets were indeed much closer to the Earth than were the stars.

Figure 6.59
This contemporary drawing shows Brahe with some of his apparatus and assistants at his Uraniborg observatory.

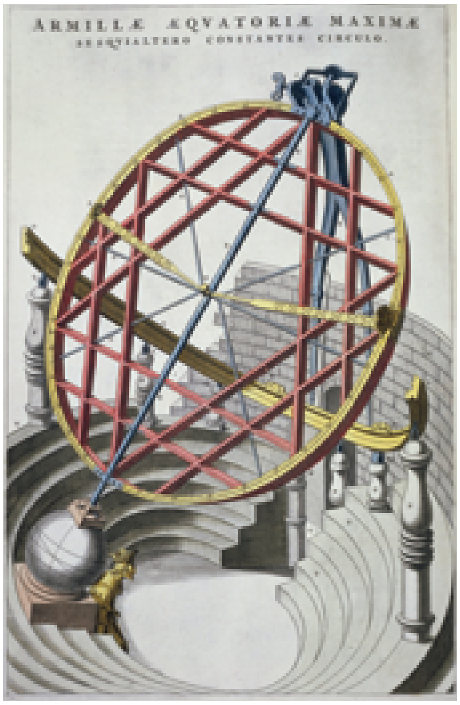


Figure 6.60

Brahe's 'Great Armillary', one of the instruments with which he measured the positions of the stars and planets.

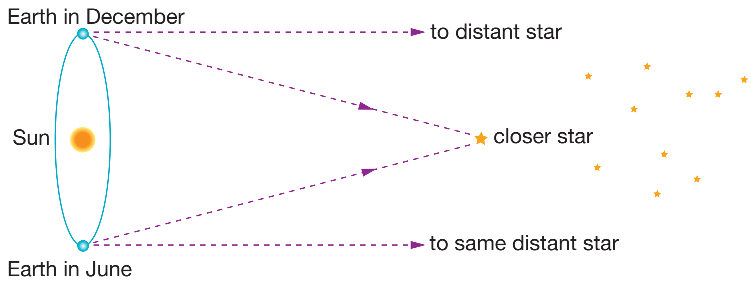


Figure 6.61

Brahe argued that if the Earth was revolving around the Sun, changing parallax should alter our view of the stars through the year. Because he found no parallax, he concluded that the Earth was at rest.

After years of painstaking observation, Brahe found no evidence of any star parallax and decided that Copernicus must have been wrong about the Earth moving. He did, however, decide that Copernicus seemed to be correct in suggesting that the other planets were revolving around the Sun rather than around the Earth. It was, he said, that the Sun, and all the planets with it, were revolving around the Earth. This was a perfectly reasonable conclusion based on his observations.

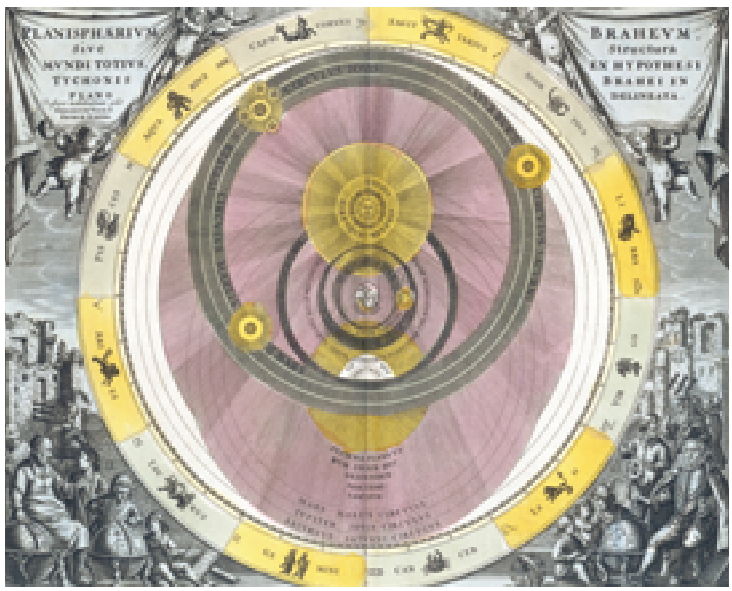


Figure 6.62

In this colourful 1660 drawing of Tycho Brahe's system, the Earth is in the centre with the Moon and the Sun circling around it, but all the planets are in orbit around the Sun. There was no way to tell the difference between this system and the Sun-centred system by observation alone.

Brahe's legacy to astronomy was the huge amount of data he had collected over the many years of his endeavours. The accuracy of his work was right at the limit of what can be achieved with observation by the naked eye and is one of the most remarkable feats of scientific observation. The real analysis of this data, however, was left to an assistant he employed in 1600, just one year before he died. That assistant was the German mathematician and keen astronomer Johannes Kepler.

Brahe and Kepler were a great team. Brahe was a painstaking observer who recorded masses of data faithfully and accurately. Kepler loved analysing that data and trying to make sense of it. Although they only worked together for the year before Brahe's death, Kepler spent much of the rest of his own lifetime analysing Brahe's data. In the meantime another great observer was using the newly invented telescope to observe the heavens. He was Galileo Galilei.

Copernicus and Brahe had opened cracks in the supposedly unchanging and perfect world of the heavens. Kepler went further by daring to question the notion that the orbits of the planets must be combinations of perfect circles. It is said that he could almost fit the orbit of Mars to a circle, but there remained a very small discrepancy with the data. However, so confident was he in the accuracy of Brahe's data that he decided the orbit could not be circular and so went on to investigate other possibilities. He finally decided that the best shape to describe the orbits of the planets was an ellipse. An ellipse can be drawn with a pencil in a loop of string around two drawing pins as shown in Figure 6.63. The positions of the two pins are the foci of the ellipse. Clearly, if the pins are moved together, the ellipse becomes a circle, and so a circle is actually a special type of ellipse. In reality most of the planet orbits look almost circular, but Kepler found that Brahe's data could only be satisfied if they were actually elliptical. We can get an idea of how close the Earth's orbit is to circular from the fact that its closest distance to the Sun (in January) is 149 million kilometres while the furthest (in July) is 151 million kilometres.

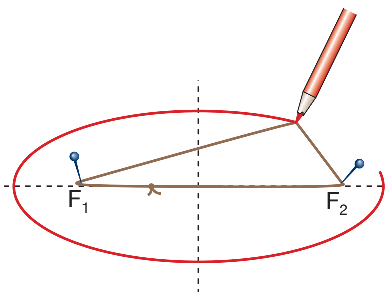


Figure 6.63
An ellipse can be drawn with a loop of string and two pins. The pins are at the foci of the ellipse.

> ***KEPLER'S FIRST LAW*** states that the orbit of a planet is an ellipse with the Sun at one focus.

Kepler found that Brahe's data fitted exactly if he assumed the orbits of the planets to be elliptical with the Sun at one focus. Each planet did not go around its ellipse at a steady speed, however: it sped up as it came nearer the Sun and slowed down as it went further away. Again, Kepler found a simple pattern in this changing speed. If a radius was drawn from the planet to the Sun, this line swept out area at a constant rate. In other words, if it swept out an area A in 30 days in January, it swept out an equal area in 30 days in July. For this to occur, the planet has to move fastest when closest to the Sun. In fact, this means that its speed along the orbit is inversely proportional to its distance from the Sun.

> ***KEPLER'S SECOND LAW*** states that the radius from the Sun to a planet sweeps out equal areas in equal times.

Having found the way to describe each planet's orbit, Kepler went on to look at the relationships between the planets. He knew that the further planets took longer to complete their orbits; that is, they had a longer 'year'. In fact a planet twice as far from the Sun as the Earth took more than twice the time of the Earth to complete its orbit, which meant that its actual speed along its orbit must be less. After considerable work (and nine years after the first two laws) Kepler discovered the relationship he was looking for.

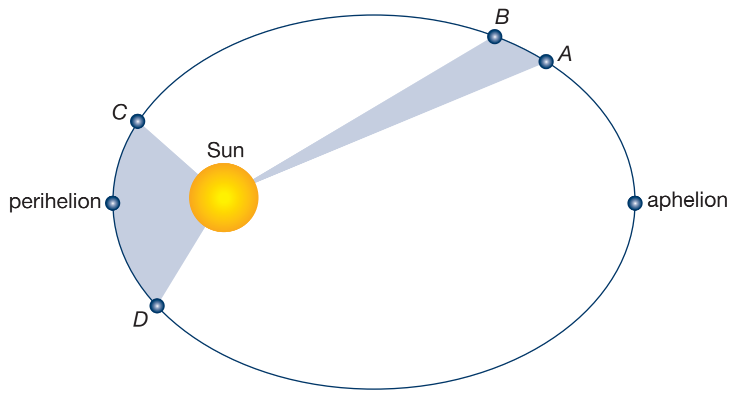


Figure 6.64

Kepler's second law says that a planet sweeps out equal areas in equal times. This means that it is moving faster at perihelion (closest to the Sun) and slowest at aphelion (furthest from the Sun). This ellipse is exaggerated; the orbits of most planets are quite close to circular.

Physics file

As the orbits are not circular, the radius of the planet's orbit varies a little. In fact the radius that is used in Kepler's third law is the semi-major axis of the ellipse. It is half the long (major) axis of the ellipse. It is close to, but not the same as, an average radius, and so is often written as '*a*' rather than *R*.

KEPLER'S THIRD LAW states that the square of the period of the planet (T^2) is proportional to the cube of the radius of the orbit (R^3). That is:

$$\frac{R^3}{T^2} = K$$

where K is Kepler's constant.

Table 6.8 illustrates Kepler's third law. If the period is measured in Earth years and the (average) radius of orbit is measured in AU, the value of K for the Earth must be 1.00 (as both R and T are 1 by definition). As you can see from the table, the values of R^3 and T^2 are equal (within experimental limits), making their ratio—Kepler's constant—equal to 1.00 (when measured in AU and years) for all planets.

table 6.8 Kepler's third law for all planets using modern values

Planet	T (Earth years)	R (AU)	T^2	R^3
Mercury	0.24	0.39	0.06	0.06
Venus	0.61	0.72	0.37	0.37
Earth	1.00	1.00	1.00	1.00
Mars	1.88	1.52	3.53	3.51
Jupiter	11.86	5.20	140.7	140.6
Saturn	29.46	9.54	867.9	868.3
Uranus	84.01	19.19	7 058	7 067
Neptune	64.79	30.06	27 160	27 160
Pluto	248.5	39.53	61 770	61 770

Not only did Kepler find a more accurate way of predicting the future positions of planets (something very important to a society still steeped in astrology) but, unlike Ptolemy, he found general patterns that all the planets obeyed. This suggested that some underlying physical principles must govern the motion of the planets. Kepler hypothesised a fairly unsatisfactory system based on magnetic forces, but it was left to two of the most well-known names in physics to discover those principles: Galileo and Newton.

6.4 SUMMARY Fundamentals of astronomy

- The stars appear to rotate as if fixed to a huge celestial sphere that is centred on the Earth. The sphere appears to rotate on an axis that passes through the Earth's North and South Poles. The axis meets the celestial sphere at the south and north celestial poles (SCP and NCP).
- The height (altitude) of the SCP in our sky is equal to our latitude. The only place from which both the NCP and SCP are visible (just) is the equator.
- The stars visible from the North and South Poles are referred to as the northern and southern stars respectively. The celestial equator is the line between them. It is the projection of the Earth's equator onto the celestial sphere.
- The position of a star on the celestial sphere is described by its right ascension (RA) and declination (dec.) coordinates. RA is the equivalent of longitude and is measured from the vernal equinox in hours, minutes and seconds. Declination is the equivalent of latitude and is measured in degrees north (+) or south (–) from the celestial equator.
- A sidereal day is the time for the Earth to rotate through one full circle; that is, the time for the stars to appear to rotate exactly once around the Earth. It is about 4 minutes less than 24 hours.
- The early Greek astronomers knew that the Earth was round, but they believed that it was fixed and in the centre of the Universe, and that everything else revolved around it.
- Ptolemy pictured the planets as undergoing epicyclic motion in which their orbits were smaller circles centred on a larger circle. This explained the retrograde motion of the planets quite well.
- Copernicus put forward an alternative Sun-centred (heliocentric) system based on the assumption that the Sun was at rest and the planets, including the Earth, circled it.
- Kepler, after analysing Tycho Brahe's data, showed that the orbits of the planets were ellipses with the Sun at one focus. He also found that the planets swept out equal areas over equal times, and that the square of the period of the planet was proportional to the cube of its radius of orbit.

6.4 Questions

Unless otherwise stated, assume that the stars are observed from within Australia, in answering these questions. The latitude of Perth is 32°.

1 As we watch the stars through the night we find that they:
  - A stay in exactly the same place in the sky all night
  - B all rise from the east and set in the west
  - C mostly rise from the east and set in the west
  - D rotate around the SCP directly overhead.

2 Briefly explain the difference, if any, you would see in the location of the SCP in the sky in Brisbane (latitude 27°) compared with its location in the Perth sky.

3 Two stars are observed from a location with latitude –30°. Star A has a declination of –30° and star B a declination of –40°.
  - a Which of the statements below about the path of star A will be true?
  - b Which of the statements below about the path of star B will be true?
    - A The star will remain overhead permanently.
    - B The star will pass directly overhead at 12 noon each day.
    - C The star will pass directly overhead at some time during the day.
    - D The star can never be seen directly overhead from that location.

4 Use a starfinder or map to find the bright stars which have the following celestial coordinates:
  - a RA 7 h 42 min, dec. +28°
  - b RA 22 h 55 min, dec. –30°

5 At 9.00 p.m. one night we see Orion due north. Explain why, when we look one week later at 9.00 p.m., it will not be due north. Approximately where will it be?

6 The right ascension of Orion is approximately 6 h and the right ascension of Aquarius approximately 23 h. Both are close to the celestial equator. If they can both be seen in the night sky, which one rose first and by how much?

7 In astronomical terms a superior planet is one that:
  - A is larger than the Earth
  - B is heavier than the Earth
  - C is further from the Sun than the Earth
  - D can support life.

8
  - a If a planet is at 'conjunction', where will it be seen in the sky? Why?
  - b If a planet can be seen high in the north sky at midnight, where is it in its orbit in relation to the Earth in its orbit? Can all planets be seen in this position?

9 Kepler discovered that the orbits of the planets were ellipses rather than circles; however, the ellipses were very close to circles. The length (long axis) and width (short axis) of the orbits for three planets are given below in AU.

Planet	Length	Width
Mercury	0.387	0.379
Earth	1.000	0.999
Mars	1.524	1.517

a What is the length expressed as a percentage of the width in each case?

b If you drew a circle of 10 cm radius to represent the short axis of the orbit, what would be the difference between this and the correct orbit?

10 a If an asteroid was found at twice the distance from the Sun as the Earth is—that is at 2.0 AU—what would be the length of its year (in terms of an Earth year)?

b If an asteroid was found which had a period of 8 years, how far would it be from the Sun?

6.5 Hubble's universe

Figure 6.65
This is NGC 2997, a spectacular gas-rich spiral galaxy. The Milky Way is thought to be somewhat similar to this galaxy.

Since the first observations with telescopes, astronomers had seen small fuzzy patches in the night sky. Many appeared irregular in shape, some round, and some even had a spiral appearance. The nature of these *nebulae*, as they became called, remained mysterious. Most astronomers thought of them as clouds of some kind relatively close by. Several astronomers, however, had suggested that at least some of the nebulae might be 'island universes'. The matter was still being hotly debated in the 1920s when a young American astronomer named Edwin Hubble found the solution to the dilemma. The problem was to determine just how far away the nebulae really were. If they were at similar distances to most stars, that would show that they were objects within our Milky Way. But they were known to be beyond stellar parallax range, and spectroscopic parallax could not be used on them.

Hubble was fortunate in that a few years earlier, Henrietta Leavitt had discovered an important relationship between the period and the brightness of a type of variable star known as *Cepheid variables*. They are named after the star Delta Cepheid, which had been discovered to vary in brightness over a period of about 5 days in 1784. In 1894, after a careful examination of the star's spectral lines, Russian astronomer Aristarkh Belopol'skii deduced that this was due to the whole star expanding and contracting at this rate. Astrophysicists now have a good model to explain this behaviour but it need not concern us here. Cepheids vary in brightness at a very regular rate, with a period of anything from a day to some weeks. Leavitt found that their average brightness was directly related to the period of the fluctuation in brightness—the longer the period, the brighter the star. The significance of this was that they could be used as distance markers because a measurement of their period enabled their intrinsic brightness to be found. By comparing this with the apparent brightness the distance could be calculated.

On careful examination of his photographs of the Andromeda 'nebula', Hubble discovered some Cepheid variables. Hubble immediately realised that as the Cepheids in Andromeda were very much dimmer than any others that had been seen, they were very much further away. His calculations put them at about 750 kpc (750 000 pc), a much greater distance than had been measured for any other star. Andromeda was no nebula; it was a whole new galaxy of stars. Based on its angular size in the sky and this estimate of its distance, it must be about 70 kpc in diameter (bigger than our own galaxy, which is less than 50 kpc). With Hubble's announcement of his discovery in 1924, whole new worlds opened up, both literally and figuratively. The Universe suddenly became even larger—and by a huge factor, as it was realised that many more 'nebulae' would turn out to be new galaxies at even greater distances. Fortunately, as you can see from the period–luminosity graph in Figure 6.66, Cepheids are very bright stars and so can be seen from a great distance. This enabled the distances to many more galaxies to be calculated.

Hubble found that there seemed to be four main types of galaxies: the spirals, the barred spirals, the ellipticals and the irregulars. The spiral and barred spiral galaxies constituted about 77% of all known galaxies, most of the others being the elliptical galaxies. The diameters of galaxies varied from a few kiloparsecs to over 200 kpc, although the diameters of the irregular galaxies were generally smaller than 10 kpc.

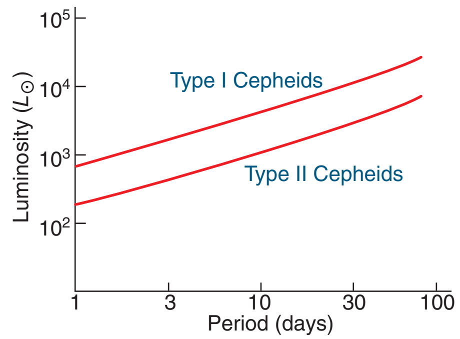


Figure 6.66
The period–luminosity relation for Cepheid variables. There are two types that can be distinguished by the metal content of their spectra. $L_{\odot}$ is the luminosity of the Sun. Fortunately Cepheids are very much brighter than our Sun, enabling them to be seen even in distant galaxies.

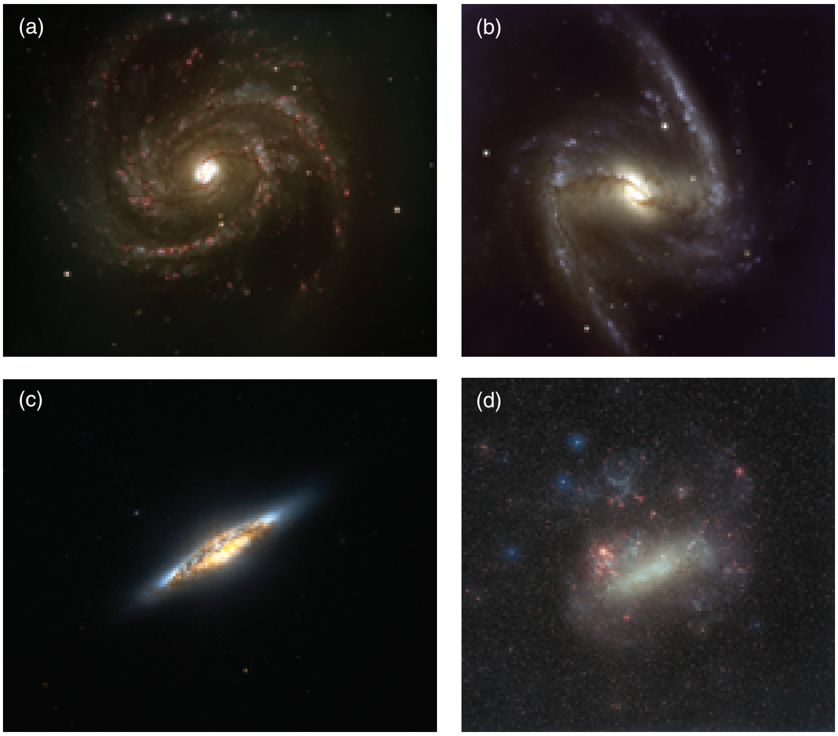


Figure 6.67
(a) NGC 4321 is a spiral galaxy with particularly pronounced arms. Many have larger central bulges and smaller arms. (b) NGC 1365 is a barred spiral because the spirals seem to originate at the ends of the bar-shaped region. (c) NGC 5010 is a typical elliptical galaxy. (d) The Large Magellanic Cloud (LMC) is a typical irregular galaxy with no obvious shape to it, as is the Small Magellanic Cloud (SMC).

The Cepheid variables visible in galaxies make them an ideal tool for distance measurements, but unfortunately they are not visible past about 30 Mpc (megaparsec). However, at greater distances astronomers have found that they can use supernovae as measuring sticks. There are different types of supernovae, but certain classes always seem to have about the same brightness. If one of these is seen in a galaxy its apparent brightness can be used to estimate its distance. This technique depends on there being the right sort of supernova explosion in a galaxy, but as time goes by more are seen.

Our own galaxy—the Milky Way

It soon became clear that the Milky Way was a galaxy like the others and of similar proportions. Instead of being populated by a vast number of stars, the Universe was suddenly populated by a vast number of galaxies. Indeed, we now know there are about as many galaxies in the Universe as there are stars in our galaxy—about 100 billion.

Physics file

Galaxies are often referred to by numbers such as M31 or NGC 224 (both of which refer to the Andromeda galaxy). The M numbers are 'nebulae' catalogued by French comet hunter Charles Messier (1730–1817). He listed them because they were sometimes confused with comets. NGC stands for the 'New General Catalogue' published in 1888 by the Armagh Observatory in Ireland. It lists over 13 000 galaxies, star clusters and nebulae. The two catalogues include both nebulae and galaxies, as the distinction was not known at that time.

Figure 6.68

We can only see the Milky Way from the inside. This is the whole-sky view of it.

Physics file

Where is the centre? Shapely's method for locating the centre of the galaxy could be likened to a motorist lost near the centre of a city. He can't see the city because of the buildings around him, but can see the light in the sky above it. Shapely assumed that there would be more stars, and in particular more clusters of stars, above and below the plane of the galaxy near its centre.

The existence of the Milky Way as a narrow band of stars across the sky suggested that our galaxy was a huge disc, rather than a sphere. As we look at the Milky Way in our night sky it appears to encircle our part of the galaxy and is fairly uniform in all directions. For this reason many people believed that the Sun was at the centre of the galaxy. However, Harvard astronomer Harlow Shapely found that a number of globular clusters of stars appeared to be above and below the plane of the galaxy in one region of the sky only. Was that the centre of the galaxy? He suggested that the reason we saw roughly equal numbers of stars in all directions was that our view in the plane of the galaxy was obscured by dust and gas that absorbed the light from more distant stars, including those in the centre of the galaxy. Shapely assumed we could see the clusters better because they were out of the plane of the galaxy. In 1920 he calculated the distance to the globular clusters and so put the centre of the galaxy at about 14 kpc (modern measurements put it at about 8 kpc).

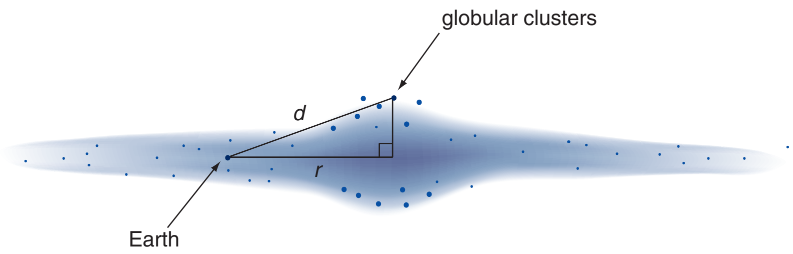


Figure 6.69

Shapely assumed that the globular clusters would be concentrated around the centre of the galaxy.

Modern astronomers, particularly those working with radio telescopes that can 'see' through the interstellar dust, have confirmed Shapely's view. We now believe the Milky Way to be disc-shaped with spiral arms, rather like Andromeda. Its diameter is about 50 kpc (although the vast majority of stars are within 15 kpc of the centre) and its thickness a little less than 1 kpc. The central bulge is about 2 kpc in diameter and located in the region of Sagittarius. However, although this is certainly a rich area of the Milky Way to observe, we cannot see the central bulge because of the obscuring effects of the increasing amount of 'dust' towards the more central area of the galaxy.

KILOPARSEC: Remember that 1 kpc = 3262 light-years.

Hubble, Doppler and the redshift

In the 1920s Hubble continued his observations of galaxies. He found that when he looked at the spectrum of a galaxy, it was 'redshifted'. This means that it had the same pattern of lines as normal spectra but they were all moved towards the red (long wavelength) end of the spectrum. This was interpreted as the well-known 'Doppler shift' that occurs whenever a source of waves is moving towards or away from us. It is most obvious as a police car goes past us. The frequency of the siren sounds higher than normal as it approaches us and then lower than normal as it recedes from us. In a similar way, if a source of light waves is travelling away from us. the frequency of the light appears to drop. As light always travels at the same speed (and $c = f\lambda$), this means that the wavelength must increase; that is, become redder. This is explained in more detail in the Physics in action on page 321. As shown there, the speed of the galaxy can be found simply from the amount of **redshift**:

$$v_{galaxy} = \frac{\Delta\lambda}{\lambda} c$$

where v_{galaxy} is the speed of the observed galaxy (m s^{-1}), $\Delta\lambda$ is the change in wavelength (m), λ is the normal wavelength (m) and c is the speed of light (m s^{-1}).

Hubble also measured the distance of the galaxies by using the Cepheid technique. He found a remarkable pattern. The more distant galaxies had greater redshifts, and indeed the redshift was directly proportional to the distance of the galaxy. In other words, the galaxies seemed to be speeding away from us, and those at the greater distances were receding at greater speeds. This became known as Hubble's law and is expressed as a simple equation:

$$v_{galaxy} = H_0 d$$

where v_{galaxy} is the recession velocity of the galaxy (km s^{-1}), H_0 is Hubble's constant (km s^{-1} Mpc^{-1}) and d is the distance of the galaxy (Mpc).

As you can see in the graph in Figure 6.70, the value of Hubble's constant is about 70 km s^{-1} Mpc^{-1}. It has proved difficult to find an accurate value for the constant because of the huge distances of the galaxies involved, but recent results from a satellite called WMAP (Wilkinson Microwave Anisotropy Probe) suggest a value of 71 km s^{-1} Mpc^{-1} to within 5%. One of the problems is that not all galaxies are receding from us. Some of the closer ones are moving towards us (and so have blueshifts) owing to their gravitational interactions. These irregular motions can mask the Hubble recession for the closer galaxies—unfortunately those for which more accurate distance measurements can be made. However, the further galaxies are all moving away at greater and greater speeds, which mask the speed of the irregular motion. It turns out that galaxies themselves come in clusters, some small, some much larger. Within the cluster, the galaxies may move around somewhat randomly, but the clusters themselves appear to be moving apart at a very regular rate—this is the Hubble constant.

Practical activity

54 Laboratory exercises in astronomy and contemporary laboratory exercises in astronomy

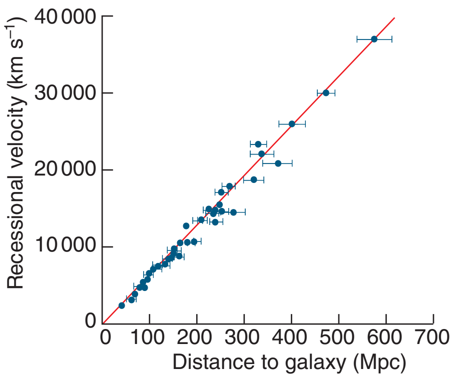


Figure 6.70

Hubble discovered that the galaxies were receding from us at velocities proportional to their distance.

The Hubble constant can be used to determine the distance to galaxies too far away to be measured by other techniques. Provided a spectrum can be seen, and thus a redshift measured, the recession speed can be determined and hence the distance. For this reason an accurate value of the Hubble constant is one of the most important quests of modern astronomy. As we will see shortly it is tied up with one of the biggest questions—how old is the Universe?

✓ Worked Example 6.5A

The spectral line of singly ionised calcium has a wavelength of 393.3 nm when measured in the laboratory, but can be seen to have a wavelength of 401.8 nm in the light from the elliptical galaxy NGC 4889.

a Find the recession speed of this galaxy.

b Use Hubble's law to find the distance to this galaxy.

Solution

a

$\lambda_{lab} = 393.3 \times 10^{-9}$ m

$\lambda_{obs} = 401.8 \times 10^{-9}$ m

$c = 3.00 \times 10^{8}$ m s^{-1}

$$v_{galaxy} = \frac{\Delta\lambda}{\lambda}c$$

$$= \frac{(401.8 \times 10^{-9}) - (393.3 \times 10^{-9})}{393.3 \times 10^{-9}}(3.00 \times 10^{8})$$

$$= 6.48 \times 10^{6} \text{ m s}^{-1}$$

$$= 6.48 \times 10^{3} \text{ km s}^{-1}$$

b $H_o = 70$ km s^{-1} Mpc^{-1}

$v_{galaxy} = 6.48 \times 10^{3}$ km s^{-1}

$$v_{galaxy} = H_o d$$

$$d = \frac{v_{galaxy}}{H_o} = \frac{6.48 \times 10^{3}}{70}$$

$$= 92.6 \text{ Mpc}$$

$$= (92.6 \times 10^{3})(3.262 \times 10^{3})$$

$$= 3.02 \times 10^{8} \text{ l.y.}$$

This distance is 302 million light-years and so we are seeing the galaxy as it was 302 million years ago.

Practical activities

16 The Doppler effect I

17 The Doppler effect II

Where is the matter?

The discovery of galactic clusters deepened a mystery already posed by galaxies themselves. Presumably it is gravity that holds all the stars to a galaxy, in much the same way that the planets are held to the Sun, and so the laws of mechanics that apply to the solar system can be also applied to galaxies. When this is done, the total mass of the galaxy can be found. The problem is that if we add up the mass of all the stars, along with the dust and gas that we can see, we get nowhere near the calculated total mass! The same process can be applied to the galaxies in a cluster, and we find the same problem. So where is the missing matter?

Physics in action — The Doppler effect

When a source of waves, whether sound, light or any other, is moving towards an observer, the apparent wavelength seen by the observer will be shorter (girl in Figure 6.71a). This is because each wave is emitted a little closer to the observer than the previous one and so is not so far behind the previous wave as it would be if the source was stationary. The reverse is the case if the source is moving away from the observer (boy in Figure 6.71a).

In Figure 6.71b two consecutive waves from a source moving with speed u are shown. The second wave has just been emitted at the position shown. The natural wavelength is λ and the period T. As the velocity of the wave is v, the wavelength $\lambda = vT$. In the time between emitting the two waves the source has moved a distance uT. This means that the distance between the waves in the forward direction is $vT - uT$, while that in the backward direction is $vT + uT$. These distances are the wavelengths as perceived by the observers to the front and rear, respectively.

In the case of light waves travelling from moving sources, we are interested in the apparent change in wavelength $\Delta\lambda$. As can be seen in Figure 6.71b, this is simply uT, so $\Delta\lambda = uT$. As T is also given by λ/v, we can rewrite this expression as:

$$\Delta\lambda = \frac{u}{v}\lambda$$

That is, the wavelength change, or 'shift' in the terminology of astrophysics (redshift or blueshift depending on the direction), is equal to the wavelength times the ratio of the speed of the source to the speed of the wave. If we replace v by the speed of light waves, c, and u by the speed of the galaxy, v_{galaxy}, and then do a little algebraic reorganisation the expression can be written as:

$$v_{galaxy} = \frac{\Delta\lambda}{\lambda}c$$

So the speed of the galaxy is equal to the fractional wavelength shift ($\frac{\Delta\lambda}{\lambda}$) times the speed of light.

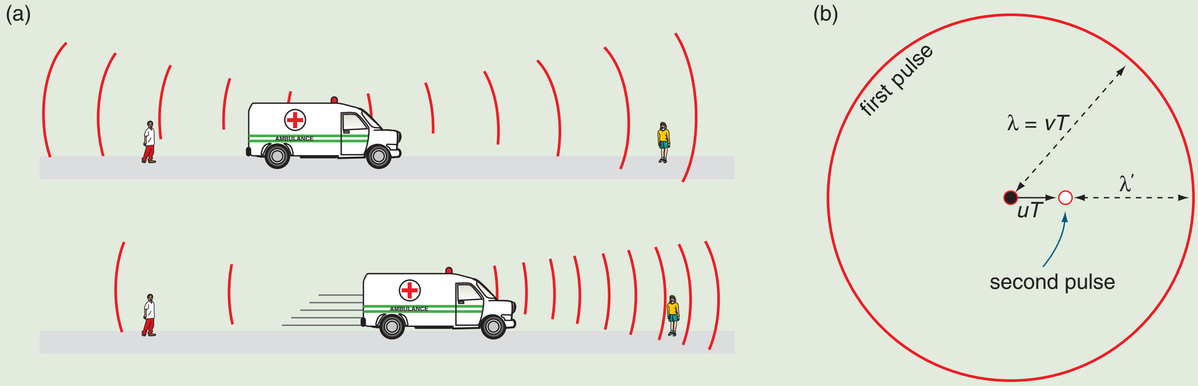


Figure 6.71
The Doppler effect.

On the way to this problem many assumptions are being made: notably, that the same laws of mechanics that apply (very successfully) to our solar system apply on the much larger scale, and that most of the matter in a galaxy is visible. Astrophysicists are reluctant to give up the first assumption and so tend to concentrate on looking for what they call *dark matter*—matter that cannot be seen through a telescope. Nevertheless, there are some indications that perhaps the laws need modification, and so some physicists are looking for solutions to the 'missing matter' problem in that direction.

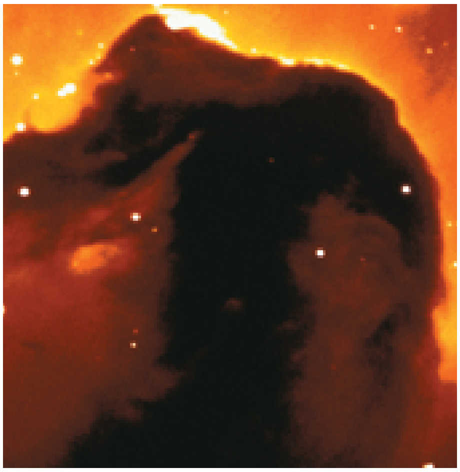


Figure 6.72
The Horsehead, a dark nebula near Alnitak, the easternmost star of Orion's belt.

Physics file

Astrophysicists often tend to use the term 'light' in the broad sense of any electromagnetic radiation. This includes radio, microwave, IR and UV as well as X-ray and gamma ray radiation. So when we speak of the apparent brightness of quasars it may not necessarily mean that they will appear to be bright objects in an optical telescope, but that the total energy we receive at all wavelengths of electromagnetic radiation is high.

It is not too surprising that all of the matter in a galaxy may not be visible. In order for us to see anything it must either emit its own light or reflect the light from another source. Some nebulae emit their own light, but many glow because of the light from nearby stars. There may well be many nebulae not illuminated in this way. Sometimes we can detect such nebulae because they block our view of the stars behind them (as in Figure 6.73). Unfortunately, however, even if estimates of this type of dark matter are included, the missing matter is still something like 90% of the total calculated mass of the galaxies. That is a lot of missing matter! At present there is no good solution to this puzzle; it is one of the ongoing mysteries of astrophysics. Various forms of dark matter have been postulated including WIMPs (weakly interacting massive particles—these are nuclear-sized particles) and MACHOs (massive compact halo objects—these are planet or larger-sized objects). Even if they have no idea what they are looking for, astrophysicists are never at a loss for creating imaginative mnemonics!

Quasars, black holes and early galaxies

Trying to find invisible dark matter somewhere out there might make looking for the needle in the haystack seem like child's play, but by now you will be used to the fact that this doesn't put astrophysicists off. To find clues to this puzzle we really need to look back in time to the early stages of the Universe. But that is just what astronomers are doing when they are looking at hugely distant galaxies. If a galaxy is at a distance of, say 2000 Mpc, which is about 6600 million light-years, in effect we are seeing it as it was 6.6 billion years ago. That is a long time ago, but astrophysicists are interested in galaxies and other objects even further away, which are even further back in time. The Hubble Space Telescope and other very large telescopes have enabled them to see objects that astrophysicists believe are actually quite 'young' in the sense that they are seeing them as they were not so long after the beginning of the Universe. For example, the Hubble telescope has photographed what are believed to be possible subgalactic gas clouds some 3400 Mpc, or 11 billion light-years, away. This means they are 'only' about 3 billion years old.

One such class of 'young' objects are the *quasars*, starlike objects that have very large redshifts and hence, it is assumed (although not entirely without question), they are at huge distances from us. One quasar, known as PC1247+3406, has a recession velocity of 94% of the speed of light, putting it at 3800 Mpc (over 12 billion l.y.). The term 'quasar' is short for quasi-stellar radio sources because they were first discovered by radio telescopes owing to the large amount of radio energy they emitted. Later, optical quasars were discovered as well, but many emit both types of radiation and so they are now all grouped together. The fascinating feature is that they seem to be so far away, and yet appear brighter than many stars in our own galaxy. This means that their intrinsic brightness is millions of times greater than that of most galaxies!

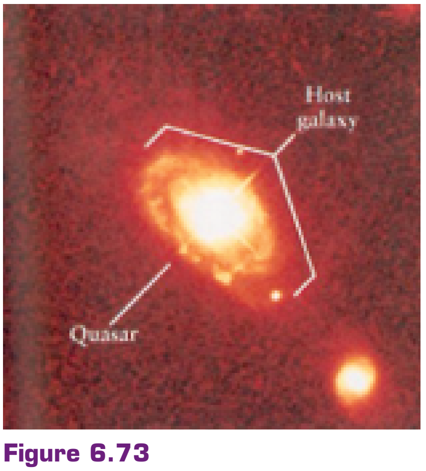


Figure 6.73

Quasar PG 0052+251 appears to be at the centre of a young galaxy. Quasars may be extremely powerful objects in the formation of galaxies.

Astronomers were puzzled that although quasars seemed quite numerous at vast distances, there were none at relatively closer distances —they would be very obvious if they were! The redshift puts the closest at about 250 Mpc. This is why many astronomers thought they might actually be objects in our own galaxy and that the redshift might be due to other unknown reasons. However, other evidence gradually convinced them that they were indeed very distant, and therefore extraordinarily bright. It appeared, then, that quasars were something that existed a long time ago (a *very* long time ago!) but not any longer. Eventually it was found that some quasars seemed to be in the centre of galaxies—but these were very difficult to see because of the brightness of the quasar. It had also been noticed (even before the discovery of quasars) that some distant galaxies had extremely bright centres. In the last couple of decades astrophysicists have concluded that quasars are somehow involved in the formation of galaxies, or more particularly *were* involved, as none seem to have survived into the last few billion years. Such powerful objects must have been extremely massive, so where did all the mass go? The answer seems to be that it ended up as *black holes* in the centre of most galaxies.

If a huge star stops producing enough energy, its gravity will cause it to collapse. We have seen that large stars end up as supernovae, which leave a superdense neutron star. The enormously powerful quasar objects would have left an even larger superdense mass. This mass would have been so dense that even the neutron matter would collapse into what astrophysicists have described as a black hole. This term basically describes a mystery! Once the density of the mass becomes so great, not even light can escape from it, hence the word 'black'. What happens inside a black hole is not known—no radiation can escape so we cannot see into it. Various theories have been put forward including such fanciful ideas as that it may be a passageway to another universe, but such thinking is pure speculation at present. A number of black holes have been found, not by seeing them directly of course, but by seeing the radiation that is given off as matter is sucked into the black hole. Because of the enormous forces involved, the matter is totally ionised and accelerated at such a rate as to give off a very characteristic type of electromagnetic radiation. It is now thought that black holes probably exist in the centre of most, if not all, galaxies.

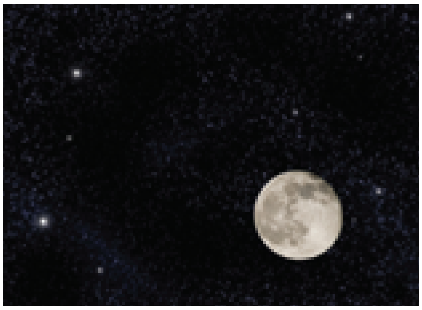


Figure 6.74
Why is the night sky dark?

The expanding universe

If all the galaxies are speeding away from each other, they must have been much closer at some time in the past. We have all heard of the expanding universe, but this is a relatively recent discovery. Before that the Universe was thought to be infinite and unchanging. At least that was the case after the time of Galileo and Newton. You may remember that Giordano Bruno was burnt at the stake in 1600 for saying that the Universe was infinite. Newton postulated that the Universe must be infinite because if it were not, in time, gravity would have pulled all the stars together into one great mass. Besides, if it was not infinite, what was beyond its edge?

There was a difficulty with this view, however. Kepler was one of the first to realise that if the Universe was infinite the night sky should not be dark. Whatever direction one looked, one should eventually see a star. Even if there was some 'dust' in the way, the dust should be illuminated from all sides. The German astronomer Heinrich Olbers brought this situation to attention again in the early 1800s and so it became known as Olbers's paradox. Even when Einstein produced his general theory of relativity in 1915, he was dismayed to find that it did not allow for an infinite universe and so amended it by adding what he called a 'cosmological constant' so that an infinite universe could exist. (Later he said this was his 'greatest blunder'—because if he had believed his own equations, it would have been he who discovered the expanding universe.)

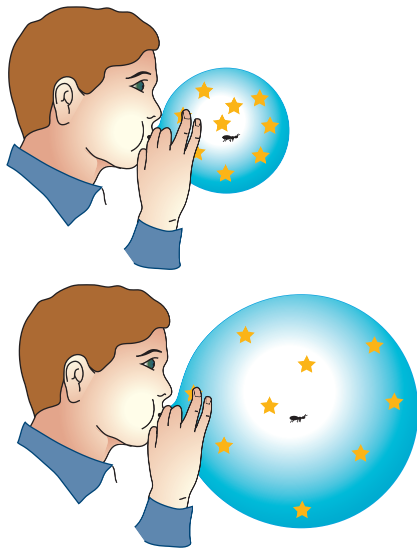


Figure 6.75

As the balloon is blown up, the stars all move away from the ant. The more distant stars will move away faster than the closer ones. In the real universe it is the galaxies that are moving away from each other.

It was Hubble who eventually resolved Olbers's paradox. He realised that because all the distant galaxies were receding from us, unless we just happened to be in the 'real' centre of the Universe, the whole universe must be expanding and therefore the light from extremely distant stars will never reach us. It is important to realise that it is not the galaxies all moving away from one point *in* space, it is *space itself* that is expanding. This is just what Einstein's equations had predicted, but even he had not wanted to believe such an outrageous idea.

A helpful analogy is to imagine yourself as an ant on a balloon that is being blown up. This balloon happens to have had little paper stars stuck onto it (not drawn on it). You would see all the paper stars moving away from you. Furthermore, wherever you wandered on the balloon's surface, you would see exactly the same thing. This is a two-dimensional analogy for what in reality is a three-dimensional situation, and so we have to bend our minds a little to imagine what this three-dimensional space would be like. One thing to realise is that just as the ant can travel in a 'straight' line and end up back where it started, so would we, if we held rigidly to what we believed was a straight course through space (and given more time than the lifetime of the Universe!) eventually end up back where we started.

The stars on our balloon were stuck on and not drawn on for a good reason. As the balloon expands, drawn stars would expand also, but the stick-on stars stay the same size. This is how astrophysicists picture the expansion of space. We don't expand, even galaxies don't expand; it is the space between them that is expanding.

It is important to make the distinction between this expansion *of* space and a conventional explosion, in which everything flies out from one point *in* space. To see uniform expansion in the explosion we would have to be right at the centre. This is not the case for the expansion of the Universe (or the ant on the balloon). Astrophysicists talk of this as the 'cosmological principle'. This states that the Universe is 'isotropic and homogeneous'. This just means that it will appear uniform in whatever direction we look and from wherever we look. Just as there is no 'centre' on the two-dimensional surface of the expanding balloon, there is no 'centre' of the three-dimensional space of the Universe. Mind boggling? Isn't astrophysics fun!

Physics file

Another analogy often used to help understand the expanding universe is that of a rising fruit loaf in which all the raisins will be seen to be moving apart from each other. Apart from the difficulty that our little ant will have in moving through the dough, the real problem is that it can come to the edge of the loaf. The whole point about expanding space is that it has no 'edge', it is not expanding *into* anything. The ant on the balloon can wander as far as it likes, it will never come to an 'edge' (given that the neck of the balloon is covered over!).

Did it all start with a big bang?

If space is expanding, then at some time in the past it must have been just a dot. Well not necessarily, said astronomer Fred Hoyle back in the 1950s. He put forward what became known as the 'steady state' theory. His idea was that the Universe really was both infinite and expanding. If it was infinite the expansion was not a problem, the 'outer' stars would never reach infinity and so could go on moving away from us forever. But wouldn't that mean that the Universe was becoming 'thinner'? And if this process had been going on forever, by now it would be ultimately thin! 'No problem', said Hoyle, 'matter is being created all the time at just the right rate to keep the density of the Universe constant'. This was not really such an outrageous idea. Quantum mechanics had already suggested that matter was less 'permanent' than we had thought. And what was the alternative: everything created in one '**big bang**'? After all, which is harder to believe?

It was Hoyle who first used the expression 'big bang' to describe a universe that started off from a tiny dot. The astrophysicists of the day really were caught in a knot. Either way they had to accept the unacceptable—that matter just came into being out of . . . what, nothing? It seemed like a really impossible task (even for astrophysicists) to try to resolve this dilemma. If Hoyle was right, where do we look for the new matter being created? If it really was a big bang, how could we possibly know? The observational evidence, an expanding universe, was consistent with both theories.

If matter was being created at a steady rate, it was not too difficult to calculate that rate from the observed density of the Universe and the expansion. It turned out that about two or three atoms of hydrogen would need to be created every day in a volume about the size of a large sports arena. There was not much point in trying to look for that! If the big bang idea was correct, what difference might it make to what we can see around us now? Not much. It seemed that only some means of looking back in time could resolve the problem. You may realise that we have already shown that this is possible, but remember that in the 1950s and 1960s there was no Hubble Space Telescope or telescope with adaptive optics.

It is important to reiterate that the big bang was not to be seen as an explosion from a small point *in* space. It was more an explosion *of space*. In fact, it was more correctly an explosion of *spacetime*. Einstein had already showed that time and space were not the separate entities we normally think of. So the big bang would not have occurred at some point in time any more than it would have occurred at a point in space. Time, space and matter were all created together.

It was well known by this time that stars had limited lifetimes, but that was not a problem for the steady state theory. New hydrogen was being created all the time, from which new stars could gradually form to replace those that died. Given that space would effectively appear to be infinitely big in either scenario, and that there did not appear to be any processes that would 'run down' the Universe with time, it seemed a daunting task to test either hypothesis. However, and not for the first time in science, the eventual answer came from an 'accidental' discovery.

In the early 1960s, Arno Penzias and Robert Wilson at the Bell Telephone Laboratories in New Jersey were trying out a new directional radio antenna designed to communicate with the new satellites that had just been placed in orbit. However, they found that wherever they pointed their antenna, they seemed to pick up some 'radio noise', microwave radiation with a wavelength of a few millimetres. After testing everything about their apparatus they could think of, they decided that the radiation was coming from outer space. Fortunately, they had heard of a new prediction made by physicists Robert Dicke and P. J. Peebles at Princeton University, just a few miles away. These physicists had calculated that in the big bang, if it occurred, the temperature near the beginning would have had to be hot enough to create a lot of helium from fusion of hydrogen nuclei. As we know, any hot object, including a hot new universe, gives off 'black body' radiation. This would have been intense high-energy, very short-wavelength gamma radiation. This radiation would have filled all of early space and simply kept radiating around ever since. So was it still present?

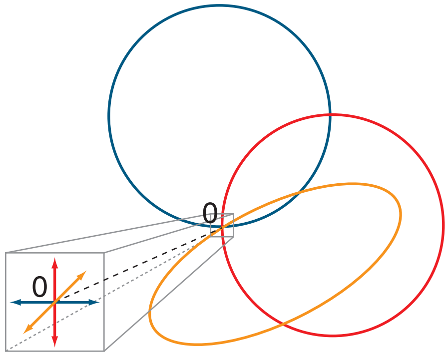


Figure 6.76
Curved space. These three circles all intersect at O. If we were to look at point O with a microscope we would see what could appear to be a set of *xyz*-axes. We have to imagine our usual view of space, as defined by *xyz* coordinates, is actually a micro-view of real space, which we can imagine as 'curved'.

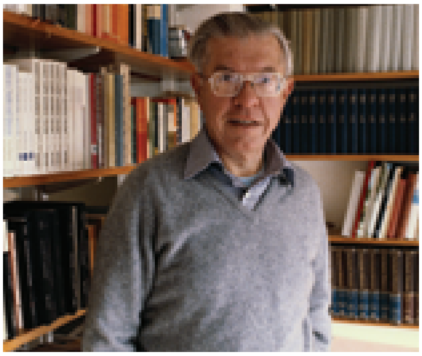


Figure 6.77
Fred Hoyle suggested that the Universe was in a steady state, with matter being continuously created to compensate for the expansion and fuel new stars. He derided the alternative, which would seem to be that the Universe must have started with a 'big bang'.

Physics file

Time, space, matter and . . . Even the laws of physics may have been created in the big bang, say some physicists. This is almost more a matter of philosophy than physics, but that is another story. It is hard not to think of the big bang as a point in time, so to avoid talking of the big bang as though it happened in time, physicists talk of it as a 'singularity'—a strange term, but it is a strange happening. Perhaps we could think of a singularity as an event that can only happen once and from which everything arises—including space and time. Asking the question 'What happened before the big bang?' is rather like asking 'What is south of the south pole?' The answer is not that there is nothing south of the south pole. There is, in fact, no answer to the question because the question itself is meaningless.

Now we know that as a gas expands, it cools. The comparison is not quite fair, but in an analogous way, as the Universe expanded, the 'temperature' of the early radiation would also decrease. (Radiation from cooler objects has a longer, or redder, wavelength.) This is because the wavelength of the radiation would expand with the space it occupied. Because of the huge expansion of space since the radiation was created, the wavelength should have increased by a huge factor up to around a millimetre, corresponding to a temperature of just a few kelvin. Now this was just the sort of radiation Penzias and Wilson had discovered! Needless to say, there were drinks all round, not to mention a great new area for research, and the death knell for Fred Hoyle's steady state theory.

While some of this **cosmic microwave background radiation**, as it is called, reaches Earth's surface, most is absorbed by the atmosphere—fortunately, or it would create a lot more radio interference. In 1989 a special satellite called the Cosmic Background Explorer (COBE) was launched. It was designed to detect and measure radiation in the range of 1 μm to 1 cm from all directions in space. If it really was the leftover radiation from the big bang it should come uniformly from all directions. COBE performed wonderfully. Its instruments showed that the spectrum of background radiation corresponded exactly to a black body with a temperature of just 2.725 K (i.e. just a little above absolute zero).

More recently another satellite, WMAP (Wilkinson Microwave Anisotropy Probe), has mapped the radiation in even more detail. While confirming the COBE results, it also measured very small variations in the temperature of the microwave background. These were only of the order of thousandths of a degree, but they were very significant to astrophysicists. If the early universe really was totally uniform, then galaxies would never have formed. In this case the gravitational pull on any piece of matter would completely balance out from all directions and so everything would just stay where it was. We would have a completely bland, and very boring universe—not good for life at all!

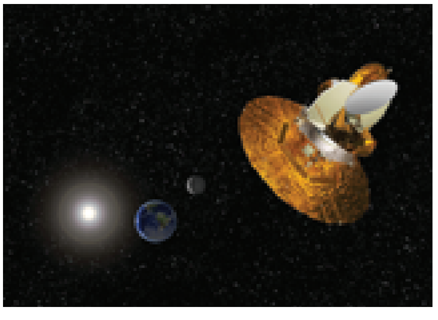


Figure 6.78

The WMAP satellite mapped the cosmic microwave background and discovered tiny fluctuations in it, which probably marked the beginning of the formation of galaxies.

In order for the galaxies to form, there had to be some anisotropy (which just means variation) in the structure of the early universe. This would enable local clumps of matter to start to coalesce and form what would become galaxies. Once a clump of matter starts to form, its gravity will accelerate the process. As the matter falls inward temperatures rise and we have the conditions for the formation of stars. However, the early anisotropy would have resulted in slight variations in the radiation produced, and hence also in its cooled-down form we see as the microwave background. This is what WMAP was looking for, and very successfully found.

When was the bang?

Given that the Universe started with a big bang, we should be able to say when it started; in other words, how old it is now. If Hubble's law is right, we simply need to run the rate of expansion backwards to see when all the galaxies were in the same place. This is not so difficult. Hubble's constant H_0 is the proportionality constant between the distance from us to a galaxy and its recessional velocity ($v_{galaxy} = H_0 d$). If the galaxy has been receding from us at a speed v_{galaxy}, and in that time has travelled a distance d, the time it has taken is $T = d/v_{galaxy}$. Now the value of d/v_{galaxy} from Hubble's equation is just $1/H_0$, so the age of the Universe is simply the reciprocal of Hubble's constant. Notice that the distance cancelled out in this expression—all galaxies, no matter how far away they are now, started off from the same point. Of course we need to convert the units into ones we recognise. We do this in Worked example 6.5B, and the answer is 14 billion years.

The age of the Universe is:

$$T_{\text{Universe}} = \frac{1}{H_0}$$

where T_{Universe} is the time the Universe has existed (s) and H_0 is Hubble's constant (70 km s^{-1} Mpc^{-1}).

✓ Worked Example 6.5B

What is the age of the Universe if the Hubble constant is taken as 70 km s^{-1} Mpc^{-1}?

Solution

The age of the Universe is the reciprocal of the constant:

$H_0 = 70$ km s^{-1} Mpc^{-1}

$$T_{\text{Universe}} = \frac{1}{H_0} = \frac{1}{70}$$

$$= 1.43 \times 10^{-2} \text{ s Mpc km}^{-1}$$

The expression for the constant includes two different units for distance, km and Mpc, so we need to convert Mpc to km. Given that 1 pc = 3.086×10^{16} m:

pc $= 3.086 \times 10^{13}$ km

Mpc $= 3.086 \times 10^{19}$ km

$$T_{\text{Universe}} = 1.43 \times 10^{-2} \text{ s Mpc km}^{-1}$$

$$= (1.43 \times 10^{-2})(3.086 \times 10^{19})$$

$$= 4.41 \times 10^{17} \text{ s}$$

$$= \frac{4.41 \times 10^{17}}{365.25 \times 24 \times 60 \times 60}$$

$$= 1.40 \times 10^{10} \text{ years}$$

That is, 14 billion years.

This method of determining the age has assumed a constant rate of expansion and ignored the effects of gravity. Neither of these is a totally valid assumption, but astrophysicists now have more complex models that do take these effects into account. Curiously enough the other effects tend to cancel each other out and the models still produce an age close to 14 billion years.

The value of 14 billion years agrees well with other estimates of the age of the Universe. The oldest stars all seem to be well under 14 billion years. Stars like our Sun are much younger. Because of the presence of heavy elements in both the Sun and Earth, we know that our solar system must have formed from dust produced by a previous supernova, so it is at least a second-generation star if not a third. Estimates of the age of the Sun put it at about 5 billion years, with the Earth and planets forming around half a billion years later.

Figure 6.79

Ellie Arroway (Jodie Foster) searches for aliens in *Contact*, the film based on the book by Carl Sagan.

It has taken about 4 billion years for intelligent life to emerge on the Earth. Given that there are about 100 billion stars in our galaxy alone, and that many of them do seem to have planets, perhaps there are many extraterrestrial beings looking up in their sky and wondering if there are other intelligent beings out there. There are various SETI (Search for Extraterrestrial Intelligence) projects looking for evidence of life 'out there'. Mostly they involve listening to likely radio frequencies that could be used to broadcast information across the galaxy. It is fascinating to wonder about the likely content of such broadcasts. Whatever we may think of such searches, as Ellie (Jodie Foster) said in the 1997 movie *Contact*, 'If there are no other intelligent beings out there, it's an awful waste of space!'

Physics in action — The future of the Universe

If the Universe is expanding, will it go on expanding forever, or will it eventually slow and perhaps start collapsing again? Naturally this question has fascinated astrophysicists ever since the idea of the big bang was suggested. If a rocket is fired upwards, it may reach a maximum height and fall back, or it may have enough energy to escape from the Earth and sail on into space forever. Or it may have *just* enough energy to escape but none left over for space adventures.

The question the astrophysicists asked was 'Did the Universe have enough energy to overcome the mutual gravitational attraction of all its mass and expand forever?' Or would gravity eventually win out and pull it all back into a 'big crunch'—and perhaps start the cycle all over again with another big bang? Actually, because of Einstein's theory of general relativity, which links space, matter and gravity, the question was put a little differently. Mass, Einstein said, distorts space. We see the effects of the 'curved space' as the effects of gravity.

In the language of relativity the question becomes whether space has positive, negative or zero curvature.

Positive curvature implies that there is enough matter in the Universe to fold it in on itself—producing a closed universe that will eventually collapse to the big crunch. Negative curvature is the opposite—space curves outwards in an open-ended way and the expansion can continue forever. In between positive and negative curvature there is the 'flat' universe—equivalent to the expansion *just* being sufficient to go on forever, but only just. These three possibilities are sometimes represented by the two-dimensional analogies shown in Figure 6.80.

There are different approaches to answering this question. One is to try to measure all the mass in the Universe and see if it is sufficient to curve space 'inwards'. The difficulty here is to detect the mass. We have already seen that most of the mass of the Universe appears to be so called 'dark matter', which by its nature is very hard to detect. Another approach

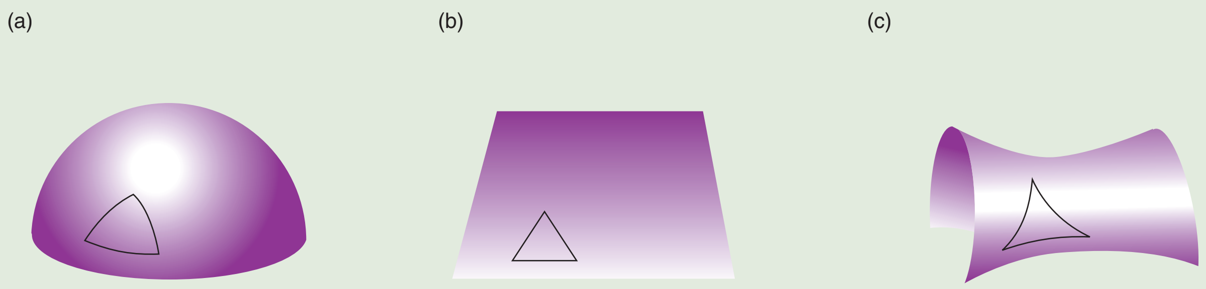


Figure 6.80

These diagrams are two-dimensional analogies for what is really three-dimensional space—or four, if time is included! In (a), positive curvature, the surface eventually folds in on itself, while the negative curvature of (c) can go on expanding forever. Surface (b) is the in-between 'flat' case.

is to measure the rate of expansion at different times in the history of the Universe to see if it is slowing down. This can be done by looking at galaxies that are vast distances from us—the further away, the further back in time we are looking.

It is the cosmic microwave background radiation (CMB) that has given us our best clue to the answer. The CMB has been travelling through space since the time of the hot early universe and has therefore been affected by the space it has travelled through. It is possible to simulate the effects of positive, zero or negative space curvature on the radiation and then compare these with the actual radiation detected by satellites such as WMAP. The differences are found in the very small irregularities we discussed earlier. The results of this research seem fairly clear—the CMB pattern is consistent with a flat universe; that is, zero curvature.

Given that the Universe is flat, the total amount of mass and energy in it can be calculated (the reverse of the process of trying to determine if it is flat from the amount of mass). Remember at this point that Einstein said mass and energy are inextricably linked ($E = mc^2$). It turns out that even with the estimated amount of dark matter, there is still a huge amount of missing 'mass-energy'. Visible matter accounts for about 4% and dark matter for about 26%. This would appear to leave 70% of the total mass-energy content missing. However, you will not be surprised to find that astrophysicists now think they know where it is!

Perhaps the most surprising results have come from the studies of very distant galaxies. They have suggested that rather than slowing down, the rate of expansion of the Universe has been accelerating. At first this seemed to be quite contrary to all expectations! Even our spacecraft with plenty of energy to escape the Earth will still slow down as gravity tugs on them. Similarly, it was thought that the mass of the Universe would eventually slow its expansion rate. However, when this acceleration was put together with the picture of a flat universe, as well as the implications of Einstein's 'cosmological constant', a new idea emerged. It seemed that the missing 70% of the mass-energy content of the Universe might be in the form of what has been called 'dark energy'. This energy, it is thought, is what is actually accelerating the expansion—it is a sort of 'anti-gravity'. As the effect of gravity (space curvature) becomes less with the increasing size of the Universe, the effect of the dark energy becomes more significant and the acceleration increases.

So how will the Universe end? Perhaps unfortunately, it does not seem that it will end with a big crunch after all. At least that may have given rise to another big bang. Rather, as everything accelerates apart it will get thinner and thinner until everything is so far away that it will become a very boring place with a very dim and uninteresting night sky—not much good for astrophysicists at all! Oh well, perhaps by then we'll have found a way to jump to one of those 'parallel universes' we hear about—one with a different, but no less interesting, set of problems for astrophysicists to solve.

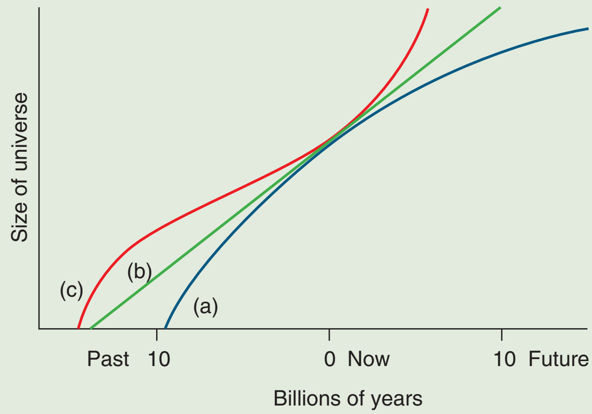


Figure 6.81

(a) This curve is for a flat universe without dark energy, and until recently seemed the most likely model. (b) This curve shows the expansion if we simply assume the Hubble constant has been constant from the beginning. (c) This curve is the latest model, which takes account of dark energy.

6.5 SUMMARY Hubble's universe

- Hubble found (using Cepheids) that certain 'nebulae' were not dust clouds but distant galaxies.
- Cepheid variables are bright stars with a predictable luminosity–period relationship and can be used to find the distance of galaxies.
- Galaxies come in four major types (spiral, barred spiral, elliptical and irregular) and tend to group in clusters. The Milky Way is spiral with a diameter of about 50 kpc and thickness less than 1 kpc. The Sun is about 8 kpc from the centre.
- Their redshift shows that galactic clusters are moving away from us at speeds proportional to their distance from us: $v_{galaxy} = H_0 d$, where H_0 is the Hubble constant.
- Galaxies have much more mass than the visible matter in them. Astrophysicists are searching for the missing dark matter.
- Quasars, extremely bright objects at vast distances, may be the beginnings of galaxies. Black holes may be the remnants of quasars.
- The implication of a stable, infinite universe was that the sky should not be dark at night. This was a problem not resolved until Hubble discovered that the Universe was not infinite but expanding.
- It is space itself which is expanding, not the objects within it. Hoyle's steady state theory explained this by postulating the continuous creation of matter. The alternative theory suggested the Universe all came into being in one 'big bang'.
- The big bang theory was confirmed by the discovery of cosmic microwave background radiation, which is radiation left over, and cooled, from the early universe.
- The age of the universe is given by the reciprocal of the Hubble constant and is found to be about 14 billion years.

6.5 Questions

1 Andromeda is a relatively close galaxy. In comparison to the diameter of our own Milky Way galaxy, how far away is Andromeda? How large would it appear in the sky if we could see it? Why can't we see it? (Use the information at the start of this section.)

2 Two similar Cepheids, which have the same period, are found in two different galaxies, but one Cepheid appears to be four times as bright as the other. What is the relative distance of the two galaxies?

3 a Given that our galaxy is about 50 kpc in diameter and 1 kpc thick, what is its approximate volume?
b If there are about 100 billion stars in the galaxy, what is the approximate volume for each star?
c From this figure, estimate the average distance between stars.
d Compare this figure with the distance of the nearest stars to the Sun and explain any apparent anomalies.

4 If a galaxy has a redshift of 10%, at what velocity is it moving relative to us?

5 The 'Coma Cluster' of galaxies is about 90 Mpc from the Earth. At what speed are they receding from us? Is this a significant fraction of the speed of light? ($c = 3 \times 10^8$ m s^{-1})

6 Why do astronomers want to look at very distant objects in the Universe? Are the distant objects different from those closer to us?

7 To account for the expansion of the Universe as observed by Hubble, two competing theories were put forward. What were they, what was the essential difference between them, and why was it seen as extremely difficult to test them?

8 What was happening in the Universe before the big bang occurred?

9 The radiation produced by the extreme conditions at the beginning of the Universe was extremely short wavelength gamma radiation. What is the relationship between this and the much longer wavelength cosmic microwave background radiation that confirmed the big bang theory?

10 Hubble's original value for the constant in his law was 500 km s^{-1} Mpc^{-1}. What would that have suggested about the age of the Universe? Do you see a problem with this age?

Chapter 6 Review

1 Name the particles thought to moderate the four fundamental forces of nature.

2 The size of the neon atom is approximately 1.02×10^{-10} m.
A Calculate the wavelength of electromagnetic radiation that will resolve the neon atom.
B Calculate the speed at which a proton would need to travel in order to resolve the neon atom.

3 Calculate the energy produced in the complete annihilation of a positron–electron pair that collide while they are both travelling at 0.500*c* towards each other.

4 List the defining characteristics of bosons, leptons, mesons and baryons.

5 What combination of quarks make a proton, a neutron and the anti-matter particles of these?

6 In 1905 Einstein put forward two postulates. Which two of the following best summarise them?
A All observers will find the speed of light to be the same.
B In the absence of a force, motion continues with constant velocity.
C There is no way to detect an absolute zero of velocity.
D Absolute velocity can only be measured relative to the aether.

7 Which of the following is closest to Einstein's first postulate?
A Light always travels at 3×10^8 m s^{-1}.
B There is no way to tell how fast you are going unless you can see what's around you.
C Velocities can only be measured relative to something else.
D Absolute velocity is that measured with respect to the Sun.

8 You are travelling from Earth towards Alpha Proxima. You notice that you are getting closer to another spaceship which remains directly in line with Alpha Proxima. Which one or more of the following could be true? This other ship:
A could be travelling towards Earth
B could be travelling towards Alpha Proxima more slowly than you
C could be stationary between Earth and Alpha Proxima
D must be heading towards Earth.

9 You are in interstellar space and know that your velocity relative to Earth is 4.00×10^6 m s^{-1} away from it. You then notice another spacecraft with a velocity, towards you, of 4.00×10^5 m s^{-1}. Which one or more of the following best describes the velocity of the other craft?
A Away from Earth at 3.60×10^6 m s^{-1}
B Towards Earth at 3.60×10^6 m s^{-1}
C Away from Earth at 4.40×10^6 m s^{-1}
D Towards Earth at 4.40×10^6 m s^{-1}

10 Spaceships A and B leave the Earth and travel towards Vega, both at a speed of 0.9*c*. Observer C back on Earth sees the crews of A and B moving in 'slow motion'. Describe how the crew of A see the crew of B, and how they see C and the Earthlings moving.
A B will appear normal, C will be sped up.
B B will appear normal, C will be slowed down.
C B will appear slowed down, C will be normal.
D B will appear sped up, C will be slowed down.
E None of these.

11 If a spaceship is travelling at 99% of the speed of light, which of the following best explains why it can not simply turn on its engine and accelerate through and beyond the speed of light, *c*, as the increase in momentum should be equal to the impulse applied?
A The law of impulse equals change of momentum does not apply at speeds close to *c*.
B While the momentum increases with the impulse, it is the mass rather than the speed that is getting greater.
C The spaceship does actually exceed *c*, but it does not appear to from another frame of reference because of length contraction of the distance it covers.
D Given enough impulse, the spaceship could exceed *c*, but no real spaceship could carry enough fuel.

12 Physicists sometimes say that the mass of an electron is about 8.00×10^{-14} J. Which of the following best explains what is meant by this statement?
A This is the 'rest energy' of the electron which, as Einstein showed, is equivalent to the mass.
B This is a misprint; it should be 8.00×10^{-14} kg.
C This is a shorthand way of saying that if all the mass of an electron was converted to energy, we would get this amount of energy.
D This is the energy of an electron which is travelling at the speed of light.

13 If you were riding in a very smooth, quiet train with the blinds drawn, how could you tell the difference between the train (i) being stopped in the station, (ii) accelerating away from the station, (iii) travelling at a constant speed?

14 What is the altitude of the celestial equator above the north horizon in Perth? Would it be different in Albany or Kununurra?

15 Use a star chart or some other reference to find the star closest to the following coordinates.
A RA 14 h 13 min, dec. +19°
B RA 5 h 50 min, dec. +7°
C RA 14 h 40 min, dec. –60°
D RA 4 h 30 min, dec. +15°

16 What are the coordinates of the following stars?
A Sirius
B Achernar
C Vega
D Rigel

17 When furthest from the Sun, the Earth is at a distance of 152.1 million kilometres and speeding around its orbit at 29.3 km s^{-1}. At closest approach the distance is 147.1 million kilometres. Use Kepler's second law to find the speed at this distance.

18 The diameter of the Sun is about 109 times that of the Earth. How many Earths would fit in the Sun?

The following information refers to questions 19 and 20.
The rate of energy output of the Sun is about 4×10^{26} W and its mass is about 2×10^{30} kg.

19 If it were made of coal, which gives out about 30 MJ per kilogram burnt, how long would it last?

20 In fact it is powered by hydrogen fusion, which releases 6×10^{14} J for every kilogram fused to form helium. How long could it last while powered this way?

21 As the stars are so far away, how can we be sure that they are not made of totally new elements that we have never seen on Earth?

22 What is it about Cepheid variables that makes them a useful tool with which to find astronomical distances?

23 Most of the 100 billion stars in the Milky Way are within about 15 kpc of the centre. Mostly the galaxy is about 1 kpc thick although there is a bulge of about 2 kpc diameter in the centre.
A What is the average distance between stars on the basis of these dimensions? Would all stars be about this far from their neighbours?
B How does this distance compare with the distance from the Sun to its nearest neighbour?

24 Astronomers depend heavily on an analysis of the spectrum of a star to determine its various properties. What are some of the features of a star that can be determined from an analysis of its spectrum?

25 The speed of sound in air is 346 m s^{-1}. If we find that the frequency of sound of a police car is 10% higher than normal, at what speed is the car moving relative to us?

26 The closest galaxies to the Milky Way are the Small and Large Magellanic Clouds. They are at distances of 50 kpc (LMC) and 63 kpc (SMC). Because they are in orbit around the Milky Way, they are not receding from us, but if they were obeying Hubble's law, at what speed would they be receding? Comment on the answer you get.

27 The Universe is said to be expanding, and yet some galaxies have a 'blueshift' when their spectrum is examined. What is the implication of a blueshift and how does this fit into the picture of an expanding universe?

28 Using the value of the Hubble constant is one way to find the age of the Universe, but there are other methods as well. What are some of these and are they consistent with the age as found from the constant?

7 Electric and magnetic fields

Australia now has the most powerful synchrotron in the southern hemisphere. The synchrotron stands on an old drive-in cinema site, next to the main campus of Monash University, Victoria. Looking something like a giant doughnut about 200 m in diameter, it produces beams of electromagnetic radiation, from infrared, through visible light, to 'hard' X-rays. The machine gives the electrons whirling within it energies of about 3 billion electronvolts and puts on a much brighter light show than the old drive-in did.

This sounds quite impressive, but what is a synchrotron? A giant waste-disposal system? An alien spacecraft? It may well look like these, but a synchrotron is actually a type of particle accelerator.

Bunches of electrons are accelerated around a huge evacuated ring to almost the speed of light. These charges are forced to follow a curved path, due to the magnetic field generated by bending magnets. As they accelerate around curves, the electrons give off bursts of radiation. This very useful radiation, called synchrotron radiation, is channelled down tubes called beamlines and utilised by researchers in a range of experimental stations.

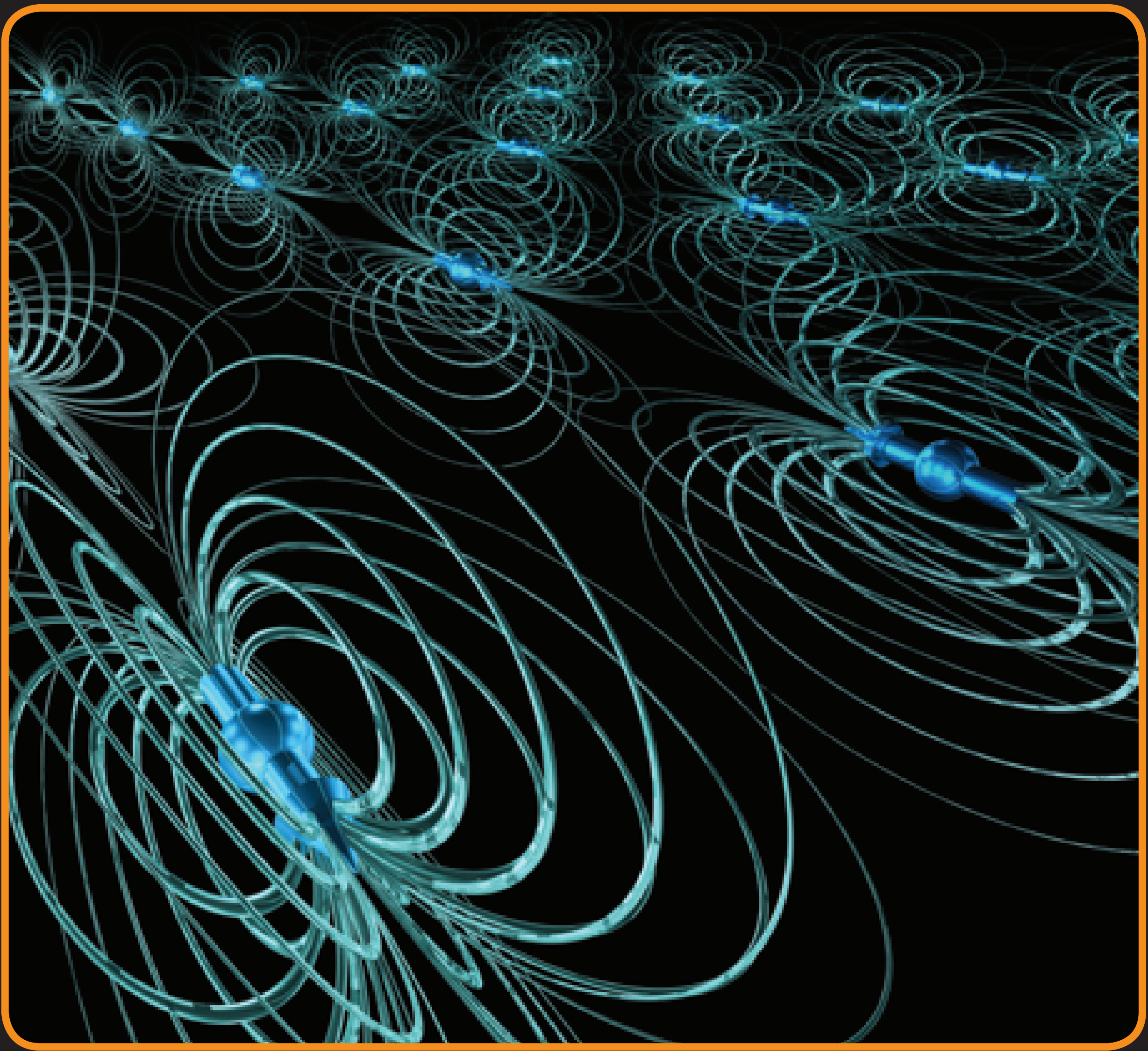
Synchrotrons can be used as super microscopes to reveal the hidden structure of fibres, chemical proteins and enzymes by using powerful techniques, the most common technique being X-ray diffraction. Synchrotrons can be used to improve medical imaging techniques, and enable us to distinguish features of cells up to 1000 times smaller than otherwise possible. X-ray lithography can be used to etch microscopic patterns on materials and construct micromachines. Experts are using synchrotron light in a broad range of experiments across a number of beamlines each day at the Australian Synchrotron.

As expressed by the 2000 Australian of the Year, immunologist Sir Gustav Nossal: 'The synchrotron will become an essential tool, helping Australian business and industry to develop new technologies and products. This project is of the highest importance to the future of Australia's capabilities in biotechnology.'

By the end of this chapter

You will have covered material from the study of electric and magnetic fields, including:

- magnetic forces on moving charges
- how oscillating electrons produce electromagnetic radiation
- acceleration of electrons in a synchrotron due to electric and magnetic fields
- the basic design of the Australian Synchrotron
- the production, characteristics and particular uses of synchrotron radiation as compared to electromagnetic radiation from other sources
- a description of types of X-ray scattering including elastic (Thomson) scattering and inelastic (Compton) scattering.
- the basic design and operation of a mass spectrometer.



7.1 Force on charges in magnetic fields

An electric current is a flow of electric charges, which might be electrons in a metal wire, electrons and mercury ions in a fluorescent tube, or cations and anions in an electrolytic cell. The nature of the flowing charge that makes up the current does not matter: the same magnetic field is produced around it, and if it is put into a magnetic field, the same force is experienced. In each case, it is simply the total rate of flow of charge—the current—that determines the field produced or the force experienced.

This suggests that the magnetic force on an electric current is really a force on the moving charges that make up that current. This is indeed the case. For example, the electrons rushing down the length of a TV tube at enormous speeds are deflected by the magnetic force they experience as they pass through the 'yoke'—the coils of copper wire at the back of the tube (Figure 7.1b).

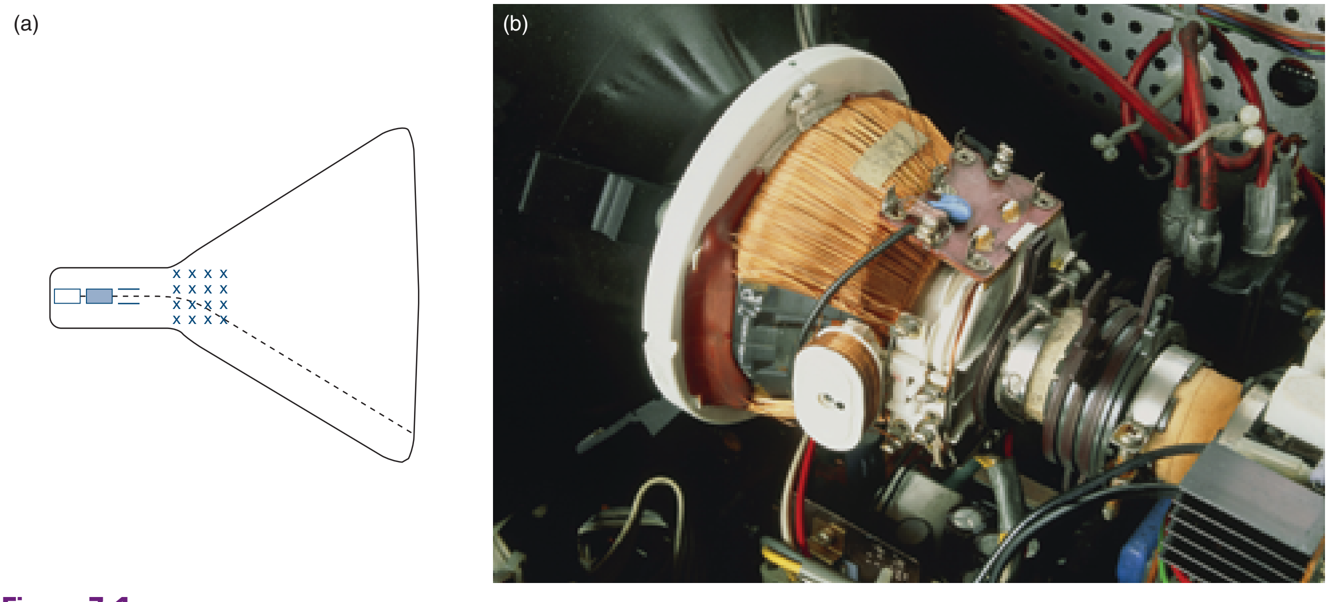


Figure 7.1
The deflection coils, or yoke, at the neck of a computer monitor cathode ray tube cause the beam to be deflected through a large angle, so the screen can be made reasonably short. The current in the coils varies rapidly in order to create the magnetic field which causes the beam to sweep across the screen at the same time as it is moving down the screen. Altogether it traces out 625 horizontal lines in two vertical sweeps of the screen, 25 times every second.

If we imagine an electric current in a wire to be equivalent to a stream of moving charged particles, it is not hard to see that a given current could be produced either by having a large number of charges moving relatively slowly or by having a smaller number of charges moving very fast. In other words the current, I, depends on the total amount of charge, Q, and the speed, v, at which the charges are moving. As the magnetic force on either of these currents is the same, the force on the faster charges must be greater to compensate for the fact that there are less of them. This suggests that the force on the individual charges that make up the current will depend on their speed, v, as well as their charge, q. Thus we should be able to find an expression for the force on an individual charge which will depend on these two quantities.

This can be done if we start with the expression for the total force on a certain length of current, and divide this by the number of charges that make up that current. The result is that the force is proportional to the charge, the velocity and the field (see adjacent Physics file).

> The magnitude of the ***ELECTROSTATIC FORCE*** ($\boldsymbol{F}_q$)
>
> $$\boldsymbol{F}_q = q\boldsymbol{v}\boldsymbol{B}_\perp$$
>
> where $\boldsymbol{F}_q$ is the electrostatic force (N), q is the charge of the particle (C), $\boldsymbol{v}$ is the velocity of the particle (m s^{-1}) and $\boldsymbol{B}_\perp$ is the perpendicular magnetic flux density (T).

Like the equation for the force on a current, this is a relationship between vector quantities. The direction of the force is as it was in the case of the force on a current, but this time the current direction is represented by the velocity of the positive charges. Remember that we have assumed the field was perpendicular to the motion of the charges. If it is not, then the component of the magnetic field that is perpendicular to the wire must be found:

If a moving object experiences a net force which is constant in magnitude and always at right angles to its motion, its direction will be changed but not its speed. As we have seen, this will result in the object moving in a circle. We can see that a particle travelling at a steady speed in a magnetic field is going to experience just this type of force and so will move with circular motion. Mass spectrometers and particle accelerators both exploit this fact. Furthermore, when the high-energy particles in the solar wind from the Sun meet the Earth's field, they experience this same force. This, it turns out, is extremely important for all life on Earth.

The Earth's magnetic field protects us from a stream of very high-energy particles, mostly protons, emitted by the Sun. As the particles approach the Earth, they encounter the magnetic field and are deflected in such a way that they spiral towards the poles—losing much of their energy and creating the wonderful auroras (the southern aurora, or aurora australis, and the northern aurora, or aurora borealis)—a visible benefit of the action of the Earth's magnetic field on moving charges!

Physics file

The force on a 1 m length of wire carrying a current I in a perpendicular field $\boldsymbol{B}_\perp$ will be given by $\boldsymbol{F} = I l \boldsymbol{B}_\perp$. If in that length of wire the current is carried by n charges, which each carry q coulombs of charge, the force $\boldsymbol{F}_q$ on each individual charge is then given by:

$$\boldsymbol{F}_q = \frac{\boldsymbol{F}}{n} = \frac{Il\boldsymbol{B}_\perp}{n}$$

If the charges are all moving at speed v, they will take $\Delta t = \frac{l}{v}$ seconds to move through this length of wire. The current will be given by $I = \frac{Q}{\Delta t}$, where Q is the total charge ($n \times q$) that moves through the length of wire in time Δt. Combining these equations gives us the current by $Il = Qv$. So the expression for the force on a single charge q becomes:

$$\boldsymbol{F}_q = \frac{Il\boldsymbol{B}_\perp}{n} = \frac{Qv\boldsymbol{B}_\perp}{n} = \frac{nqv\boldsymbol{B}_\perp}{n} = q\boldsymbol{v}\boldsymbol{B}_\perp$$

Physics file

The right-hand palm rule is used again to find the direction of the force on the moving charges. The fingers represent the direction of the perpendicular component of the field, $\boldsymbol{B}_\perp$, the thumb this time represents the direction of the velocity, $\boldsymbol{v}$, of *positive* charge q, and again the force on the charges is in the direction you would normally push with your right hand. The direction of the force on a negative charge, for example an electron, is just the opposite.

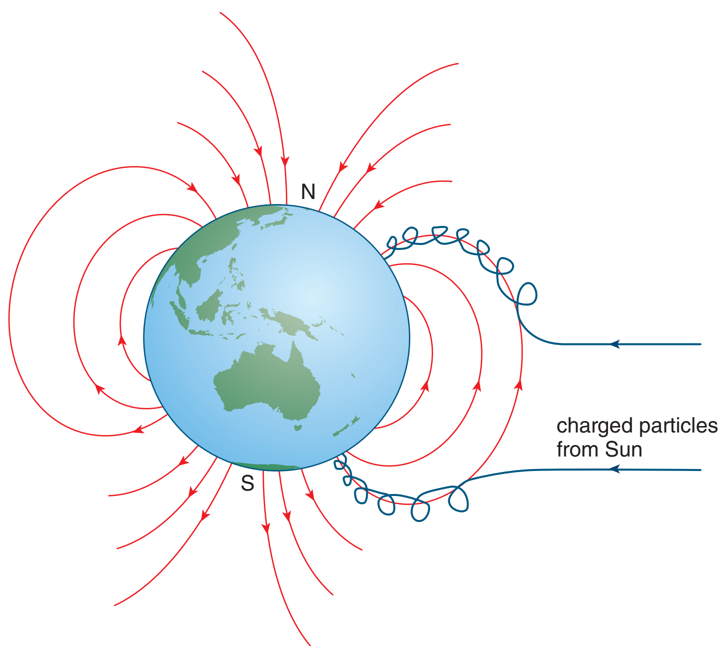


Figure 7.2
Charged particles from the Sun or deep space are trapped by the Earth's magnetic field, which causes them to spiral towards the poles. As they do this, they lose their energy and become harmless to the life on Earth.

It seems reasonable that the magnetic force on a moving charge depends both on its charge and the strength of the field. The fact that it also depends on its velocity is a reminder that this is what magnetism is really all about—a force that depends on the motion of electric charges. It is worth noting that the field, $\boldsymbol{B}_{\perp}$, itself is also due to moving charges: either an electric current or the electrons spinning in their atomic orbits. Fundamentally, magnetism is a force between charges moving relative to each other.

For more information see Chapter 5 'Wave properties and light'.

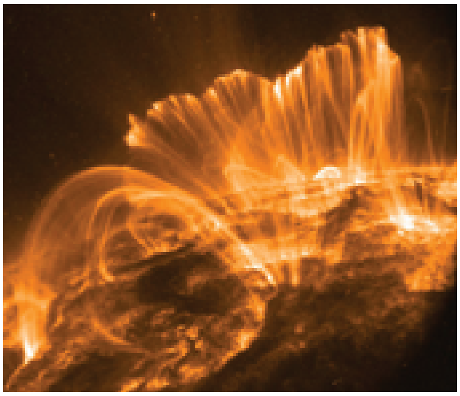


Figure 7.3

Solar flares are caused by huge ejections of gas from the Sun's surface. Because the gas is so hot it is ionised and so traces out a gigantic arc as the charged particles move in the Sun's enormous magnetic field. This flare is over 20 times the size of the Earth.

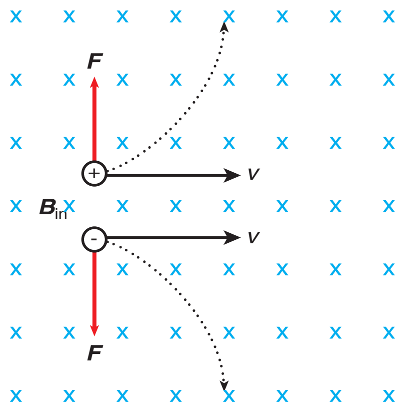


Figure 7.4

The direction of the force on a positive and a negative moving charge as they travel in a magnetic field directed into the page. Use the right-hand palm rule to confirm these directions for yourself.

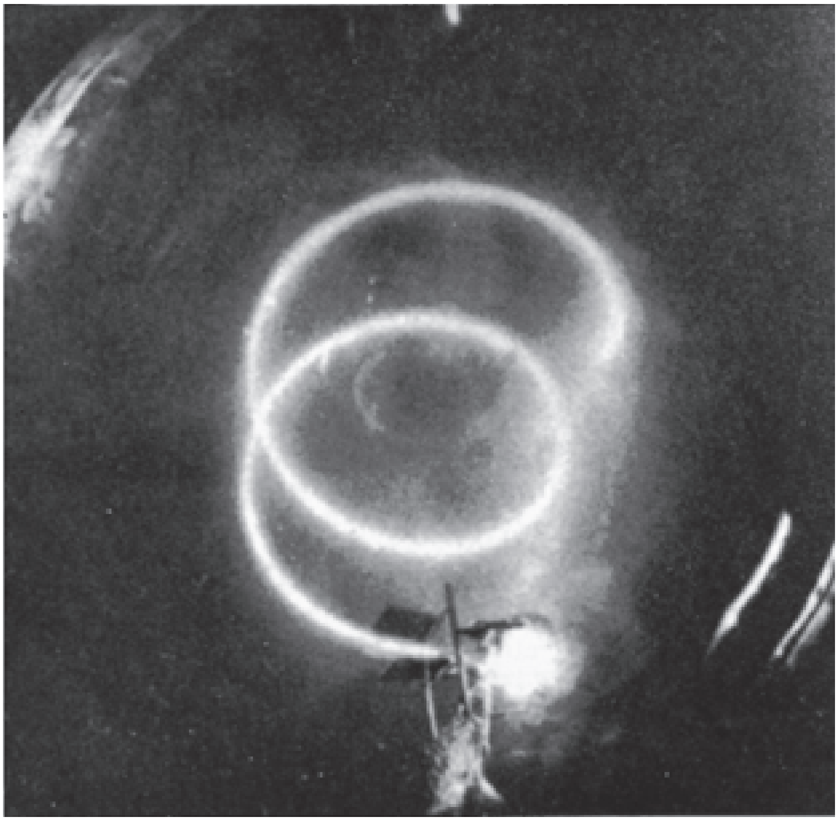


Figure 7.5

A fluorescence produced by electrons moving through gas at a very low pressure in a glass bulb. The electrons are being fired from an electron gun (similar to that used in a TV tube) into a magnetic field which is more or less perpendicular to the page. Can you tell whether the field is pointing into or out of the page?

Physics in action — Mass spectrometers and particle accelerators

One of the most powerful tools of the modern chemist is the mass spectrometer. It enables the very accurate determination of the mass of the atoms in a tiny sample of material. This can help chemists to identify the substances in very small samples, for example tiny amounts of a rare element in a sample of Moon rock, or the mud on the shoes of a suspected criminal.

The sample is heated in an evacuated container until it is ionised, and the positive ions are accelerated electrically to high speeds. They then enter a magnetic field where they experience a 'qvB' force which will cause them to move in a circle. The radius of this circle depends on the mass of the particle: heavier particles with the same charge and speed will experience the same force, but because of their greater inertia they will move in a larger circle. Thus ions of different masses end up in a different place on the detector. As the relative amounts of each ion can also be measured, the composition of the original sample can be determined. (A new type of mass spectrometer, called an *ion trap*, does not use this principle at all. It uses high-frequency radio waves to trap ions of a certain mass.)

To study the basic constituents of matter, physicists accelerate particles (such as electrons or protons) to very high speeds and then let them crash into atoms. The results of these collisions have revealed that an incredible array of subatomic particles exists. Understanding the properties of these particles is fundamental to understanding our Universe.

The electrons or protons are accelerated by electric fields of various sorts, but very long paths are needed to obtain the extremely high speeds necessary (close to the speed of light). Because it is impractical to have a straight path hundreds of kilometres long, the particles travel through very strong magnets which cause them to move in a circle. The circular accelerator at CERN in Switzerland is nearly 9 km in diameter! Closer to home, the new synchrotron near Monash University in Melbourne is 70 m in diameter. It can be much smaller because it accelerates electrons, not protons, and it uses more powerful magnets to bend the electrons into the circular path.

The Australian Synchrotron accelerates electrons to a massive 3 GeV of energy. This is equivalent to accelerating them through 3000 million volts. At this energy, they travel at about 99.99999% of the speed of light. Because of relativistic effects, their effective mass at that speed is about 6000 times the rest mass. The path of the electrons is not so much a circle, but a series of straight sections in between 24 bending magnets 2 m long and with field strengths of about 1.5 T. As the electrons go through these magnets, because they are being accelerated, they give off energy in the form of electromagnetic radiation—light. As well as the bending magnets, there are other magnets called undulators and wigglers which, as the names imply, accelerate the electrons back and forth, again causing them to give off enormously intense light. It is this light, ranging from infrared and visible to hard X-rays, that is used for many different research purposes.

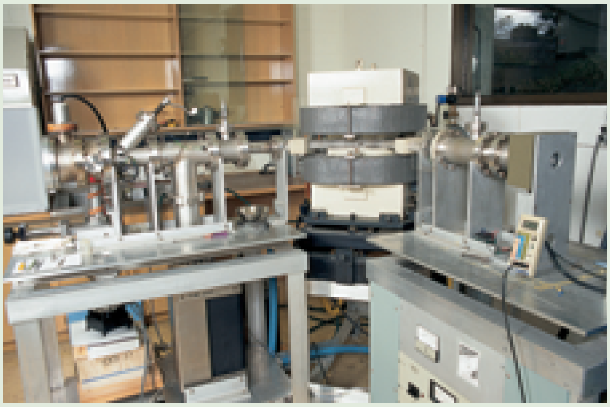


Figure 7.6
A simple mass spectrometer

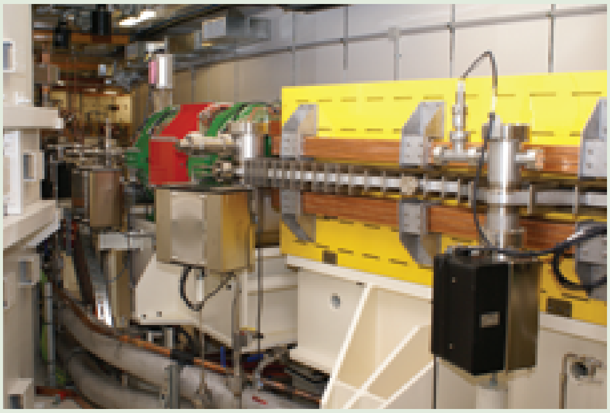


Figure 7.7
The Australian Synchrotron

7.1 SUMMARY Force on charges in magnetic fields

- Fundamentally, the magnetic force on an electric current is due to the force on the moving charges.
- The magnitude of the force $\boldsymbol{F}_q$ on a charge q moving at speed v in a field $\boldsymbol{B}_\perp$ is given by $\boldsymbol{F}_q = qv\boldsymbol{B}_\perp$. The direction is at right angles to the motion and to the field.
- Charges moving freely at right angles to a magnetic field will move in a circular path.

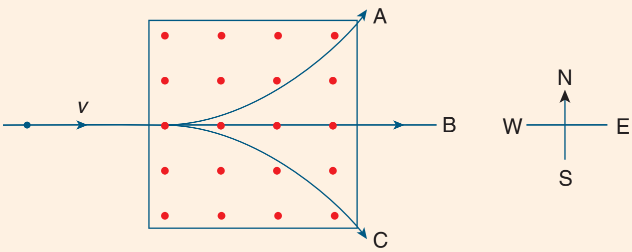
7.1 Questions

1 Which of the following quantities does not affect the magnitude of the magnetic force experienced by a particle moving through a magnetic field?
  A charge
  B velocity
  C mass
  D field strength

2 Which one of the following is correct? The force on a charged particle moving through a magnetic field will be a maximum when its velocity is:
  A perpendicular to the direction of the field
  B parallel to the direction of the field
  C at an angle θ to the field that is between 0 and 90°.

The following information applies to questions 3–6.

The following diagram shows a particle, with initial velocity ***v***, about to enter a uniform magnetic field, ***B***, directed out of the page.



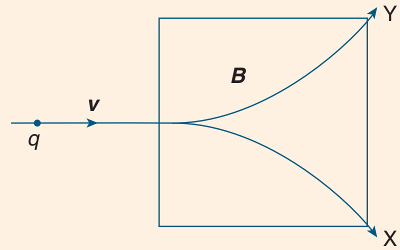
3 If this particle is positive:
  a what is the direction of the force on this particle just as it enters the field?
  b which of the paths A, B or C will this particle follow?

4 Which of the following statements is correct?
  A The kinetic energy of this particle remains constant in the field.
  B The momentum of the particle remains constant in the field.
  C The acceleration of the particle is zero in this field.

5 If this particle is negatively charged:
  a what is the direction of the force on the particle just as it enters this field?
  b which path will the particle follow?

6 What type of particle could follow path B? Explain why.

7 The following diagram shows a particle with charge *q* entering a uniform magnetic field of strength ***B***. The initial velocity, ***v***, of the particle is perpendicular to the field. Which of the following combinations would result in the particle following path X?

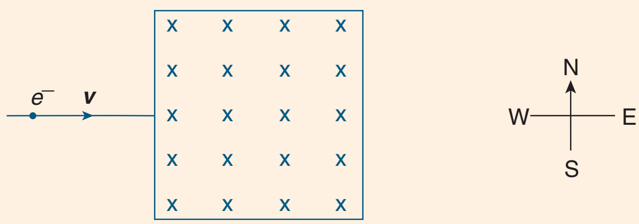


  A *q* is positive, ***B*** is into the page.
  B *q* is negative, ***B*** is out of the page.
  C *q* is positive, ***B*** is out of the page.
  D *q* is negative, ***B*** is into the page.

8 Which of the alternatives in Question 7 would result in the particle following path Y? Explain why.

The following information applies to questions 9 and 10.

The following diagram shows an electron, whose charge is *e*, about to enter a uniform magnetic field, ***B***, directed into the page. The initial velocity, ***v***, of the electron is perpendicular to the field. The magnitude of the magnetic force on this electron in the field ***B*** is ***F***.



9 a What is the direction of the magnetic force ***F*** on the electron just as it enters the field?
  b In terms of *F*, what would be the magnitude of the magnetic force if:
    i the velocity is doubled and field strength remains the same?
    ii the velocity remains the same, but the field strength is doubled?
    iii both the velocity and field strength are doubled?

10 An alpha particle (charge +2*e*, with mass about 7500 times that of the electron) enters the field with a velocity ***v***.
  a In terms of ***F***, what would be the magnitude and direction of the force on this alpha particle just as it enters the field?
  b How would the curvature of its path compare to that of the electron?

7.2 Particle accelerators

A synchrotron is a type of particle accelerator. Particle accelerators are machines that were originally designed to investigate the nature of matter by examining the structure of atoms and molecules. Charged particles, such as electrons, protons or atomic nuclei, are accelerated to speeds often close to that of light. These particles travel through an electric field, inside a hollow tube pumped to an ultra-high vacuum, with pressures comparable to those found in deep space. Strong magnets direct the particles to collide with a target or with another moving particle. Scientists obtain information about the make-up of the subatomic particles fired from the machine, or the target samples that are hit, by analysing the types of collisions that occur.

One of the first particle accelerators was the Van de Graaff accelerator, similar to the Van de Graaff generator. Developed in the 1930s, it can accelerate charged particles between metal electrodes to energies of about 15 MeV before they collide into a fixed target. Currently, the world's most powerful particle accelerator is located at the Fermi National Accelerator Laboratory (Fermilab) in Illinois, USA. It can produce energies of 1 TeV. Two sets of particles can be accelerated in opposite directions around its central evacuated ring, to meet in a collision of mammoth energies!

In contrast to these types of particle accelerators that were built for collisions, a synchrotron light source is designed to use electrons to generate beams of infrared, UV, visible and X-ray radiation, rather than colliding the electrons themselves into a target. It is this radiation, called synchrotron light, that is then channelled off for use in experimental stations to analyse materials.

Figure 7.8
This tandem Van de Graaff accelerator uses two generators to produce beams of charged particles that are accelerated by potential differences of up to 10 million volts.

Cathode ray tubes

A cathode ray tube is a useful type of particle accelerator. Electrons are released from a negative terminal, or hot cathode, in a vacuum, and accelerate towards a positive terminal, or anode. The beam of electrons is collimated, or narrowed as it passes through a slit, and releases light when it hits a fluorescent screen. A potential difference of around 2–3 kV exists between the cathode and the anode, which causes the charged particles to accelerate. Older style televisions (excluding plasma and LCD screens), visual display units and cathode ray oscilloscopes (CROs) all consist of cathode ray tubes.

The electron gun

A computer monitor, cathode ray oscilloscope or larger scale particle accelerator relies on a source of charged particles to be accelerated. The device used to provide these particles is called an *electron gun*.

Electrons are, in effect, boiled off a heated wire filament, or cathode. They are accelerated from rest across an evacuated chamber towards a positively charged plate, or anode, due to the electric field created between charged plates. Once the electrons continue through a gap in this positive plate, their motion can be further controlled by additional electric and magnetic fields. Focusing magnets are also used to control the width of the beam.

In 1875, Sir William Crookes developed a number of tubes to study cathode rays. The type of cathode ray tube shown in Figure 7.12 is called

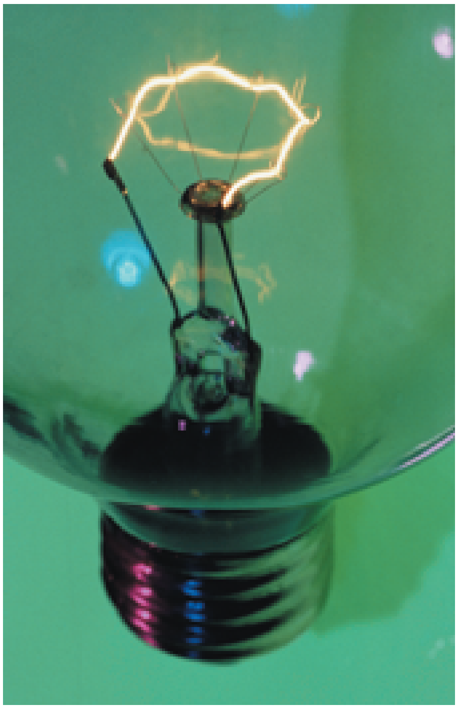


Figure 7.9
Electrons are accelerated from the heated filament within an incandescent light bulb which produces photons of light.

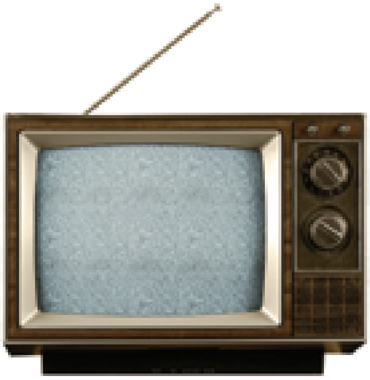


Figure 7.10
The picture tube of these older style televisions each act as a type of particle accelerator, using large voltages to accelerate electrons along the tube.

a Maltese cross tube. Inside this, electrons are accelerated from the hot cathode towards the anode by a high potential difference. The cathode and anode are contained within an evacuated glass bulb that is coated with fluorescent material. The electrons travel in straight lines and cast a dark shadow of the Maltese cross against the blue or green fluorescent background. A magnet brought near the tube can be shown to deflect the electrons and even make the shadow disappear.

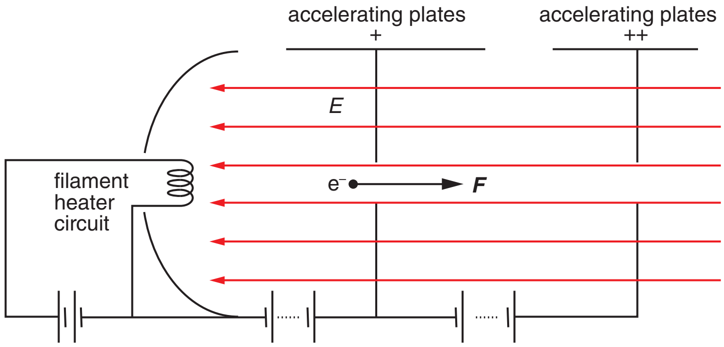


Figure 7.11
The set-up of a typical electron gun assembly. Electrons are released from a hot cathode and are accelerated across a potential difference through slits in a pair of positively charged plates. A magnetic field helps to centralise the path of the electrons.

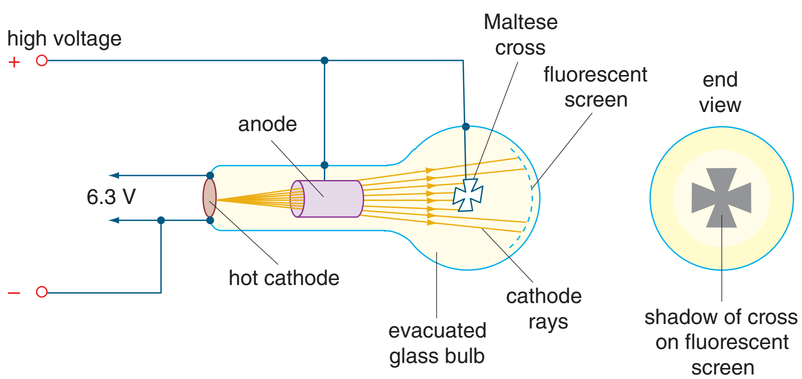


Figure 7.12
The Maltese cross tube.

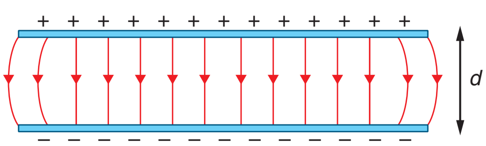


Figure 7.13
The electric field between a pair of oppositely charged plates. The direction of the field is defined as being the direction in which a positive charge would move in this region.

Accelerating electrons using an electric field

Consider an electric field acting on an electron as the result of a pair of oppositely charged parallel plates connected to a DC power supply. The electron is attracted to the positive plate and repelled from the negative plate. An electric field is acting upon any charged particle within this region. This electric field is a vector quantity and may be compared in some ways to the Earth's gravitational field. This electric field has units N C^{-1} and is defined as:

$$\boldsymbol{E} = \frac{\boldsymbol{F}}{q}$$

where $\boldsymbol{F}$ is the force (N), $\boldsymbol{E}$ is the electric field (N C^{-1}) and q is the electric charge (C).

A charge will then experience a force equal to qE when placed within such an electric field.

Recall that the magnitude of the electric field may also be expressed as:

$$\boldsymbol{E} = \frac{\Delta V}{d}$$

where $\boldsymbol{E}$ is the electric field (N C^{-1}), d is the separation of the plates (m) and ΔV is the potential difference (V).

We can substitute this value for electric field strength to produce an expression for the force on a charge within a pair of parallel charged plates to be:

$$\boldsymbol{F} = \frac{q\Delta V}{d}$$

In addition, we can consider the energy gained by an electron as it is accelerated towards a charged plate by the electric field. The work done in this case is equivalent to:

$$W_d = q\Delta V$$

We can use this equation to calculate the increase in kinetic energy as an electron accelerates from one plate to another. If a charge is accelerated from rest from an electron gun, then:

$$\begin{aligned}\Delta E_\text{k} &= W_\text{d} \\ &= q\Delta V\end{aligned}$$

The effect of a charged particle in a magnetic field

To explore the forces acting on a beam of electrons in a particle accelerator, we also need to consider the effect of a magnetic field on a charged particle. From your work in Chapters 3 and 4, you will remember that because an electric current is itself a stream of moving charges, we can state that:

$$\boldsymbol{F} = q\boldsymbol{v}\boldsymbol{B}_\perp$$

where $\boldsymbol{F}$ is the force (N), q is the charge (C), $\boldsymbol{v}$ is the velocity (m s^{-1}) perpendicular to a magnetic field and $\boldsymbol{B}_\perp$ is the perpendicular magnetic flux density (T).

So, in the case of the magnetic force on an electron moving within the magnetic field of a particle accelerator:

$$\boldsymbol{F} = e\boldsymbol{v}\boldsymbol{B}_\perp$$

The direction of the magnetic force exerted on the charge is predicted by the right-hand palm rule. Note that the direction of current is defined as the direction in which a positive charge would move, so this direction must be reversed to correctly predict the direction of motion of an electron.

If the magnetic field is not perpendicular to the motion of the electrons, then the component of the magnetic flux density that is perpendicular to the direction of the electron's velocity must be found.

If a moving charge experiences a force of constant magnitude that remains at right angles to its motion, its direction will be changed but not its speed. In this way, bending magnets within a particle accelerator act to alter the path of the electron beam, rather than speed the electrons up. As a result, the electrons will follow a curved path of radius r.

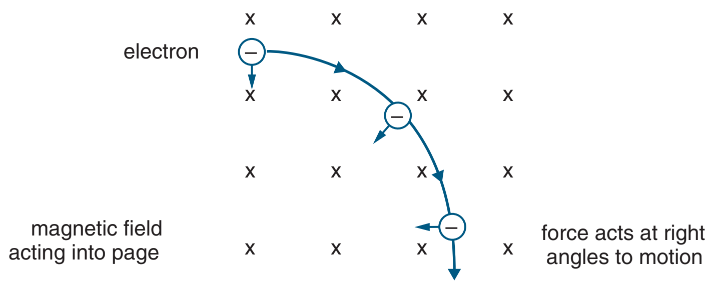


Figure 7.14

An electron fired horizontally to the right into a magnetic field that acts into the page will initially experience a force vertically downwards as predicted by the right-hand palm rule. It will follow a curved path of radius *r*. Note that the direction as defined for the current in the right-hand palm rule must be reversed to provide the direction that an electron will follow.

In this case we can say that the net force acting on the charge is:

$$\boldsymbol{F} = m\boldsymbol{a}$$

This is equivalent to the magnetic force on the charge, so that:

$$m\boldsymbol{a} = e\boldsymbol{v}\boldsymbol{B}_{\perp}$$

The acceleration in this situation is centripetal and has magnitude:

$$\boldsymbol{a} = \frac{v^2}{r}$$

Substituting this value into the previous equation:

$$\frac{mv^2}{r} = e\boldsymbol{v}\boldsymbol{B}_{\perp}$$

Rearranging this equation, we can find an expression to predict the radius of the path of an electron travelling at right angles to a constant magnetic field as:

$$r = \frac{mv}{e\boldsymbol{B}_{\perp}}$$

As the momentum of the charged particle is equivalent to $m\boldsymbol{v}$, we can also state this radius as:

$$r = \frac{\boldsymbol{p}}{e\boldsymbol{B}_{\perp}}$$

where $\boldsymbol{p}$ is the momentum of the electron (kg m s^{-1}).

Physics file

You will recall that the British scientist J. J. Thomson used an apparatus in 1897 to deflect electrons using electric and magnetic fields and then calculated the charge-to-mass ratio for electrons. The magnetic and electric fields were at right angles to each other and exerting opposing forces on the electrons within the cathode ray tube. By adjusting the strengths of the electric and magnetic fields, Thomson was able to find the point of no deflection of the electron path. At this stage, the magnetic and electric forces on the electrons are equal. When Thomson switched off the electric field, the resulting path of the electrons was due solely to the magnetic forces acting. A curved path was described which was due to the centripetal force from the magnetic field. Through knowing values for the electric and magnetic field strengths and measuring the radius of the electron path. Thomson calculated charge-to-mass ratio to be 1.759×10^{11} C kg^{-1}. He knew that the electron must be a subatomic particle as this ratio was some 1800 times smaller than the charge-to-mass ratio of the hydrogen ion. Robert Millikan extended this work in 1909 to produce values for the charge and mass of the electron.

✓ Worked Example 7.2A

An electron gun releases electrons from its cathode, which are accelerated across a potential difference of 32.0 kV, a distance of 30.0 cm between a pair of charged parallel plates. (Assume that the mass of an electron is 9.11×10^{-31} kg, the charge on an electron is -1.60×10^{-19} C and ignore any relativistic effects in your calculations.)

a Calculate the strength of the electric field acting on the electron beam.

b Calculate the magnitude of the velocity of the electrons as they exit the electron gun assembly.

c The electrons then travel through a uniform magnetic field perpendicular to their motion. Given that this field is of strength 0.200 T, calculate the expected radius of the path of the electron beam.

Solution

a $\Delta V = 32.0\times10^{3}$ V

$d = 30.0\times10^{-2}$ m

$$E = \frac{\Delta V}{d} = \frac{32.0\times10^{3}}{30.0\times10^{-2}} = 1.07\times10^{5}\ \text{V m}^{-1}$$

b $\Delta V = 32.0\times10^{3}$ V

$e = 1.60\times10^{-19}$ C

$m = 9.11\times10^{-31}$ kg

$$E_k = W_d$$
$$\tfrac{1}{2}mv^2 = e\Delta V$$
$$v = \sqrt{\frac{2e\Delta V}{m}} = \sqrt{\frac{2(1.60\times10^{-19})(32.0\times10^{3})}{9.11\times10^{-31}}} = 1.06\times10^{8}\ \text{m s}^{-1}$$

This velocity is approximately 35% of the speed of light.

c $B_{\perp} = 0.200$ T

$e = 1.60\times10^{-19}$ C

$m = 9.11\times10^{-31}$ kg

$v = 1.06\times10^{8}$ m s^{-1}

$$r = \frac{mv}{eB_{\perp}} = \frac{(9.11\times10^{-31})(1.06\times10^{8})}{(1.60\times10^{-19})(0.200)} = 3.02\times10^{-3}\ \text{m}$$

This means the electrons will follow a path of radius 3.02 mm. In actual fact, this extremely small radius is not realistic, due to the effects of relativity.

Linear accelerators

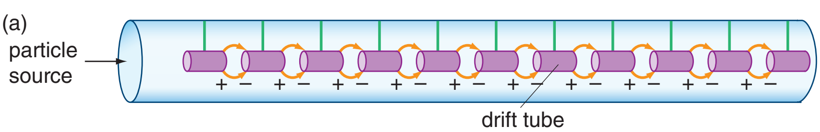


Figure 7.15
(a) Charged particles entering the linac from the left are accelerated towards the drift tubes by an electric field. They travel through the drift tubes at constant velocity, before being accelerated again in the gaps between each tube and the next. (b) This is the standing wave linear accelerator found at the Berkeley Laboratory. Can you see the drift tubes?

Cathode ray tubes are useful particle accelerators but are limited to using voltages of a few tens of kilovolts. A linear accelerator, or linac, accelerates particles in straight lines. The first linac was built in 1928 by the Norwegian engineer Rolf Wideröe. It consisted of three hollow metal tubes inside an evacuated cylinder. These are called drift tubes and were used in Wideröe's machine to accelerate potassium ions to an energy of 50 000 eV (50 keV).

The type of linear accelerator that is also called a standing-wave linear accelerator consists of a large number of drift tubes, each separated from the next by a gap. Electrons enter the cylinder and are accelerated towards the first drift tube by an electric field. An alternating potential difference is applied to each tube. This is timed so that the tube is positive as electrons approach it and negative as they exit. In this way, they are accelerated across each gap between the drift tubes. Inside the drift tube, they travel at a constant velocity because they are shielded from the effects of the electric field. The particles pick up more energy every time they leave the drift tubes, until they are accelerated out of the linac.

Physics file

Stanford University is home to the world's longest linear accelerator. It is 3.2 km long and can accelerate electrons to energies of 50 GeV. This is a travelling-wave type of linear accelerator. The machine was designed to cause two beams of particles to collide so that researchers could explore the make-up of fundamental particles of matter, such as weak bosons.

This linac consists of electric fields set up as standing waves throughout the evacuated cavity. It is useful for low-energy ion accelerators (less than 200 MeV) and for non-relativistic particles.

The type of linac used to accelerate electrons, such as that employed in the Australian Synchrotron, makes use of travelling, as distinct from standing, waves. This is called a travelling-wave linear accelerator and is explored in the next section. Linear accelerators have a number of uses, including the production of X-rays to treat deep tumours.

Physics in action — Cyclotrons

To perform experiments at very high energies, linear accelerators would need to be extremely long. For this reason, the American physicist Ernest O. Lawrence designed the first circular accelerator in the 1930s. This is called a *cyclotron*, and it won Lawrence a Nobel Prize in 1939. In some respects, the cyclotron operates as a spiral-shaped linac. Protons are often used as the accelerating particles in this machine.

Here, the many drift tubes are replaced by two semicircular, D-shaped, hollow copper chambers, called dees. These are the positive and negative electrodes of the cyclotron between which exists a strong electric field. The dees sit back to back, giving the cyclotron its circular shape, and lie between the poles of a powerful electromagnet. The inside of the metallic dee is shielded from the electric field. The magnetic field acts on the particles, producing a circular path. When a particle emerges from the dee, the sign of the accelerating potential is reversed, so the particle speeds up towards the other dee. This occurs so that a proton will accelerate towards a negatively charged dee as it exits the positively charged dee. Each time the particles cross the gap between the dees, their speed increases and they travel in a semicircle of larger radius. They gain energy with each revolution until they attain sufficient energy to exit the accelerator.

A key to the operation of the cyclotron is that the frequency of the radio-frequency generator (RF generator) that produces the alternating field must match the frequency of the circulating charged particles. The charged particles travel in a path of radius:

$$r = \frac{mv}{q\boldsymbol{B}_\perp}$$

Their speed is then:

$$v = \frac{rq\boldsymbol{B}_\perp}{m}$$

and the time taken for one orbit of the cyclotron is:

$$T = \frac{d}{v} = \frac{2\pi r}{v}$$

Substituting for r in the above expression:

$$T = \frac{2\pi r}{v} = \frac{2\pi mv}{vq\boldsymbol{B}_\perp} = \frac{2\pi m}{q\boldsymbol{B}_\perp}$$

Strangely enough, the time taken for one revolution of the cyclotron does not depend upon the velocity of circulating charges. This is because as the speed increases, the radius of path travelled also increases and the time taken for each orbit remains the same.

In 1943, the Adelaide-born physicist Marcus Oliphant, while working in Britain, suggested modifying the cyclotron design to produce a synchrotron.

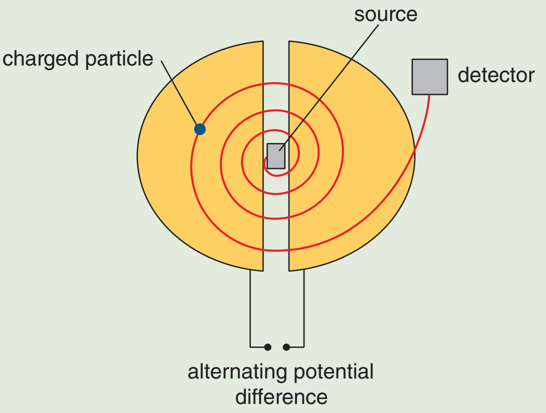


Figure 7.16
The cyclotron operates in some ways like a spiral linac—particles are accelerated from a source, through semicircular chambers called dees, until they gain sufficient energy to exit.

Physics in action — CERN

CERN is the European Organisation for Nuclear Research. Founded in 1954, it is the world's largest particle accelerator research centre, and it was developed to explore the make-up of matter and the forces that exist in the Universe. It is an international collaboration, currently with 20 member states, and is located on the border of France and Switzerland, just outside Geneva.

The accelerator complex consists of a number of separate accelerators. The large electron positron collider (LEP) was an enormous machine that began operating in 1989. It was the largest particle collider in the world, built with a 27 km circumference ring which was located some 100 m underground. This particle accelerator ceased operation at the end of 2001 and a new collider, called the large hadron collider (LHC), was constructed within the original LEP ring. Commencing operation in 2008, the LHC was designed to search for phenomena that have been predicted by theoretical physics. It is capable of colliding two counter-rotating beams of protons with energies of 7 TeV (1 TeV = 1000 GeV) each, resulting in collisions of some 14 TeV in energy! Some 1740 superconducting magnets positioned around the ring will guide the particle beam on its journey. These magnets must be cooled to −271°C, which is colder than deep space! The magnets are housed within a cryogenic distribution line to achieve such low temperatures.

Figure 7.17
This photograph shows the gentle lowering of a 1900 kg section of the CMS (compact muon solenoid) being assembled as part of the LHC. Bunches of protons are expected to collide up to 40 million times every second within this detector. Physicists hope to create the rare Higgs boson particle in such experiments.

Figure 7.18
Part of the old tunnel of the LEP collider, now refitted in the construction of the LHC. Around 6500 scientists use the facilities at CERN, including half of the world's particle physicists. The Internet itself originated as a result of physicists at CERN being interested in developing a method of sharing information across long distances.

Beams rotate around the ring for several hours at a time and the collisions take place within the four main dedicated experimental stations set up at the facility. Physicists hope that by studying the effects of high-energy collisions, they will unlock more of the secrets about the smallest particles in the Universe.

Physics file

Hundreds of cyclotrons are used worldwide to produce about 20% of the medical radioisotopes injected into patients. These cyclotrons typically accelerate protons to energies of around 40 MeV. The protons are bombarded into the nuclei of target atoms, which become proton rich and thus unstable. They then usually become stable through a process of radioactive decay. The radioisotopes produced vary in half-life up to about 3 days. Table 7.1 is a sample of radioisotopes produced in cyclotrons and their uses.

table 7.1 Some radioisotopes and their uses

Radioisotope	Use
Rubidium-81	Diagnostic imaging tool for the lung
Iodine-123	Implants to treat stenosis (abnormal narrowing) of body cavities
Thallium-201	Diagnosis of coronary artery disease and various heart conditions
Gallium-67	Tumour imaging

7.2 SUMMARY Particle accelerators

- In a particle accelerator, charged particles are accelerated, sometimes to speeds close to that of light.
- A cathode ray tube is a simple particle accelerator in which particles accelerate from a hot cathode to a high-voltage anode.
- An electron gun is the source of electrons in a cathode ray tube and many particle accelerators.
- An electron may be accelerated across a potential difference, such as in the case of an electron accelerating towards an anode in a cathode ray tube.
- An electron within an electric field will experience a force equivalent to $\boldsymbol{F} = q\boldsymbol{E}$, where $\boldsymbol{E}$ is the electric field strength. This can be rewritten as $F = \frac{q\Delta V}{d}$ as the electric field strength is equivalent to $E = \frac{\Delta V}{d}$, where ΔV is the potential difference between plates and d is the plate separation.
- The work done on an electron of charge q accelerating across this potential difference V is $W_d = q\Delta V$.
- The increase in the kinetic energy of an electron of mass m with final velocity $\boldsymbol{v}$ is:
$$\Delta E_k = q\Delta V = \frac{1}{2}mv^2$$
- For the case of an electron moving at right angles to a magnetic field, the force it experiences is:
$$\boldsymbol{F} = qv\boldsymbol{B}_\perp$$
If the magnetic force is the net force acting on the electron, then it will move in a circular path of radius r:
$$r = \frac{mv}{q\boldsymbol{B}_\perp}$$
- A linear accelerator, or linac, accelerates particles in straight lines. The standing-wave linac is generally used to accelerate low-energy ions and the travelling-wave linac, such as that used in the Australian Synchrotron, is employed to accelerate electrons.

7.2 Questions

For each of the following questions, assume that the mass of an electron is 9.11×10^{-31} kg and its charge is 1.60×10^{-19} C.

1 Electrons in a cathode ray tube are released from a:
A hot anode
B hot cathode
C cool cathode
D cool anode.

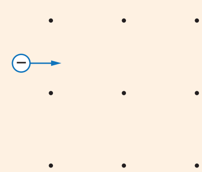
2 Sir William Crookes developed a number of tubes to study cathode rays.
a Describe the basic set-up of these cathode ray tubes.
b Describe how the electrons are accelerated through the tubes.

3 Describe the operation of the standing-wave linac. Why is this type of linac not employed in the Australian Synchrotron?

4 A cyclotron may be described as a spiral-shaped linac. What is the major advantage in having particles accelerated in a circular rather than linear fashion?

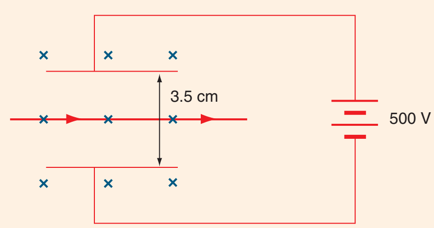
5 An electron gun assembly emits electrons with energies of 10 keV. Ignore the effect of relativity in answering the following questions.
a Calculate the magnitude of the predicted exit velocity of the electrons.
b Upon exiting the electron gun assembly, the electrons now enter a uniform magnetic field of 1.50 T acting perpendicular to their motion. Calculate the predicted radius of the electron beam.

6 This diagram represents an electron being fired at right angles towards a uniform magnetic field acting out of the page.



a Copy the diagram and mark on it the continued path you expect the electron will follow.
b Which factors would alter the path radius of the electron as it travels?

7 A stream of electrons travels in a straight line through a uniform magnetic field and between a pair of charged parallel plates, as shown in the diagram.



Calculate the:
a electric field strength between the plates
b speed of the electrons, given that the magnetic field is of flux density 1.50×10^{-3} T.

8 Electrons in a cathode ray tube are accelerated through a potential difference from a cathode to a screen. Calculate the speed at which they hit the screen if the potential difference between electrodes is 2.50 kV.

9 An electron with speed of 7.60×10^6 m s^{-1} travels through a uniform magnetic field and follows a circular path of diameter 9.20×10^{-2} m. Calculate the magnetic flux density of the field through which the electron travels.

10 a Calculate the force exerted on an electron travelling at speed of 7.00×10^6 m s^{-1} at right angles to a uniform magnetic field of strength 8.60×10^{-3} T.

b Given that this force directs the electron in a circular path, calculate the radius of its orbit.

7.3 Synchrotrons

Synchrotron light was first discovered in the 1940s when it was observed being produced in particle accelerators used for theoretical physics. When first discovered, this radiation was seen as an unwanted by-product of the acceleration process, as its release robbed accelerating particles of energy. It was only later that the useful benefits of such radiation became apparent. The synchrotron machine was designed to have a constant path radius for the particle beam, which meant that a ring or doughnut-shaped layout could be used. Since their origins in the 1940s, synchrotrons have undergone progressive evolution.

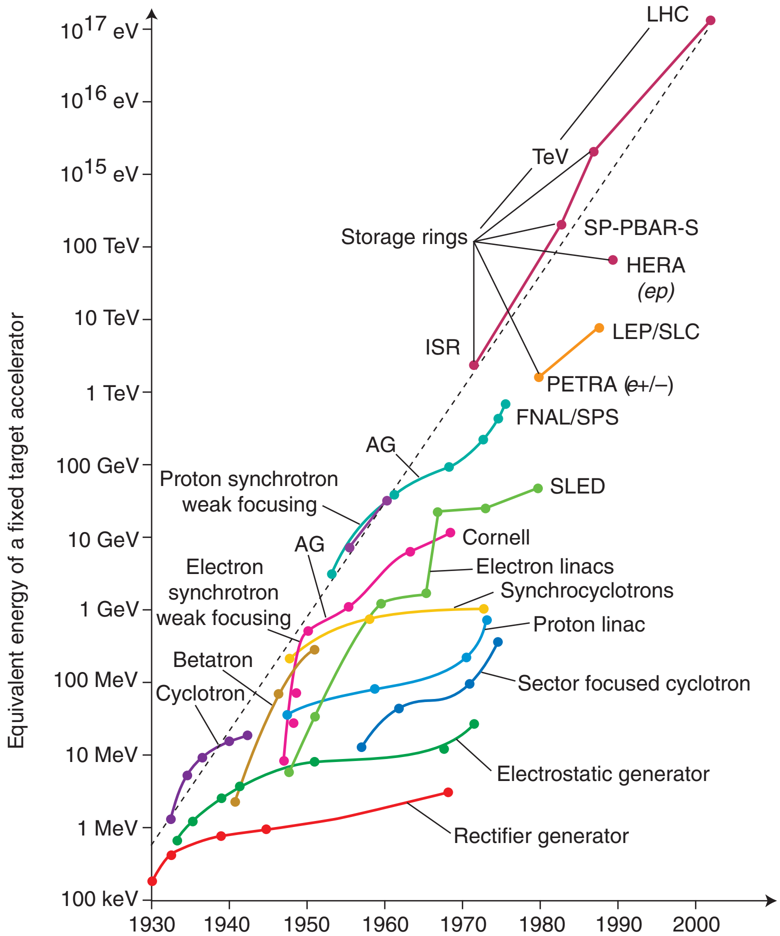


Figure 7.19

This graph demonstrates the rapid rise in capabilities of particle accelerators dating from the 1930s to the present day.

How do they work?

Figure 7.20 features the major components of synchrotron design: the linear accelerator (linac), the booster ring, the storage ring, beamlines and experimental stations. These components are outlined in the following sections.

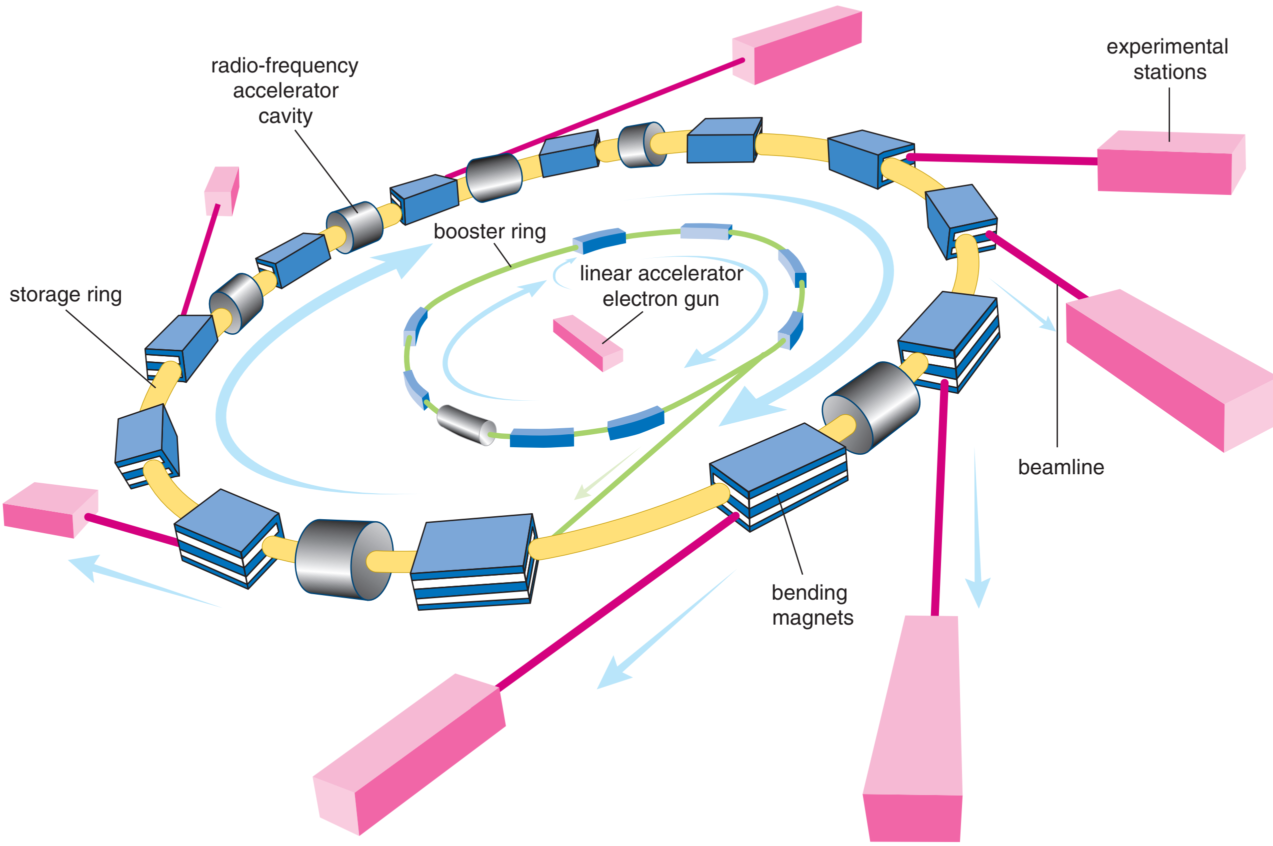


Figure 7.20
The Australian Synchrotron is almost the size of the MCG oval. This scale of some 67 m in diameter is necessary to contain the electrons, which are travelling at almost the speed of light as they zoom around the storage ring. So, how does it actually work?

The linac

We have already considered how a standing-wave type of linac reliant upon drift tubes operates. The Australian Synchrotron and other electron linacs make use of travelling waves rather than standing waves in order to accelerate particles. The travelling-wave linac consists of an electron gun, a vacuum system, focusing elements and RF (radio-frequency) cavities.

Electrons escape from the electron gun as they boil off the heated filament of the assembly. From here, they accelerate across a potential difference of about 100 keV and exit the electron gun at a velocity of approximately half the speed of light. At such velocities, the effects of relativity must be considered, as the mass of the electrons is actually greater than their rest mass. Although a full consideration of the effects of relativity lies outside the scope of this course, some impacts will be discussed more fully in the next section.

The electron beam travels through an ultra-high vacuum within the linac, to prevent energy loss through interaction with air particles. As the electrons travel, focusing elements act on the beam to constrict it to a narrow beam in the centre of the vacuum tube.

Electrons are accelerated to close to the speed of light after their journey through the linac. Such acceleration is critical to the production

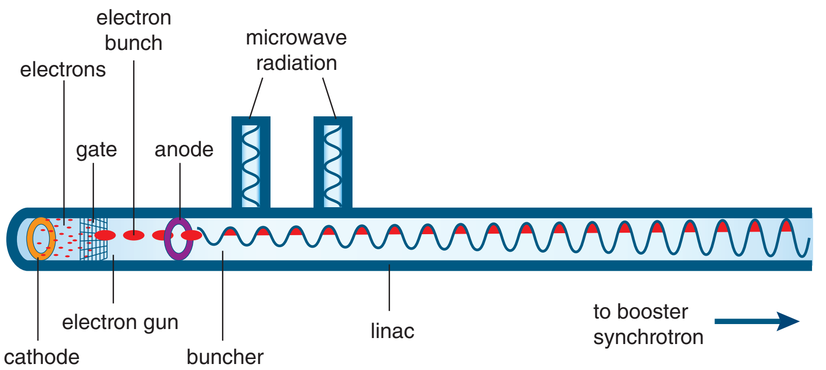


Figure 7.21
A basic travelling-wave linac design, showing the electrons being emitted from the electron gun, accelerating towards an anode and then being further accelerated in bunches by RF radiation.

of synchrotron light in the storage ring. You may wonder how such acceleration is achieved.

The answer lies in the function of the cylindrical RF (radio-frequency) cavities (also known as klystrons) that surround the electron beam. These cavities produce intense electromagnetic radiation at several hundred megahertz. The RF radiation propagates through the linac as a travelling wave. When timed correctly, electrons can, in effect, 'ride the crest' of this RF wave, resulting in their acceleration to enormous speeds.

In the Australian Synchrotron, electrons are released from the gate of the electron gun in pulses every 2 ns to travel towards the anode. These electrons accelerate as they pass through the RF cavity with the crest of the RF radiation and are slowed down when passing in conjunction with the trough of the RF radiation. This effect causes the electrons to become bunched into groups as they travel through the linac itself. The frequency of RF radiation is timed to accelerate the arrival of each electron bunch. The linac used in the Australian Synchrotron gives the electrons a kinetic energy of 100 MeV.

The booster ring

Within the circular booster ring, bending magnets provide a force at right angles to the motion of the electrons in order to bend them into a circular path. In this ring, the energy of the electrons is increased (boosted) from 100 to 3000 MeV, or 3 GeV (gigaelectronvolts). The energy boost is supplied by a radio-frequency (RF) chamber through which the electrons travel on each orbit of the ring.

If we were to simply calculate the speed of the electrons at these energies using the equation $\frac{1}{2}mv^2 = eV$, we would find that the speed in the booster ring would have gone from about 6×10^9 to 3×10^{10} m s^{-1}. However, these speeds are well over the speed of light! We all know that Einstein discovered that nothing could exceed the speed of light ($c = 3.00 \times 10^8$ m s^{-1})—so what does happen?

Einstein's very famous equation $E = mc^2$ shows us that somehow energy and mass are interrelated. This is the explanation of the apparently strange results of our simple calculation above. As we keep adding energy to the electrons we find that, once we are near c, it is not the speed so much as the mass which increases. Although the equation $E_k = \frac{1}{2}mv^2$ breaks down at relativistic speeds (at, say, speeds above about 10% of c) it reminds us that kinetic energy depends on mass as well as the speed. Normally we assume (quite reasonably) that m remains constant, but in Einstein's relativity we find that this is not the case at very high speeds.

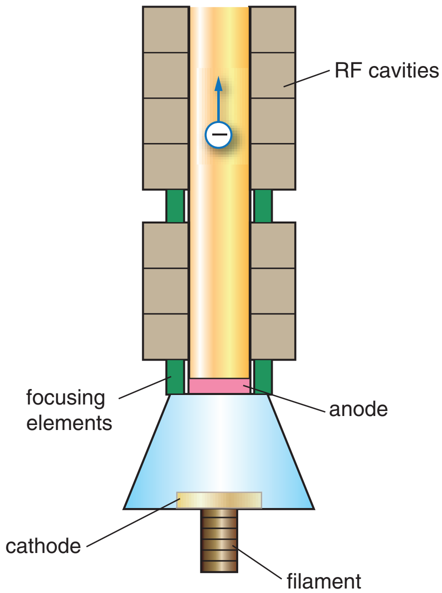


Figure 7.22
In the Australian Synchrotron, electrons boil off the heated filament of the electron gun assembly. The gate in this diagram is a grid that is also a cathode. A bias voltage prevents the electrons from travelling towards the anode. A 500 MHz RF voltage is applied every 2 ns to overcome the bias voltage, allowing electrons to travel in pulses towards the anode.

Figure 7.23
Looking like a mass of metal components, this photograph shows part of the RF system of the Australian Synchrotron.

Einstein found that the effective mass of an object increases very sharply as its speed approaches *c*. Even at 10% of *c*, the mass has only increased by about 0.5%, at 87% it has doubled, and at 99% of *c* the mass has increased by about seven times. After 99% of *c*, the mass increases very rapidly. In the 3 GeV Australian Synchrotron, the speed of the electrons is about 99.99999% of *c* and the mass is about 6000 times the rest mass. So the 3 GeV of energy pumped into each electron shows up, not as a huge speed increase far greater than *c*, but as a 6000-fold increase in the mass of the electron.

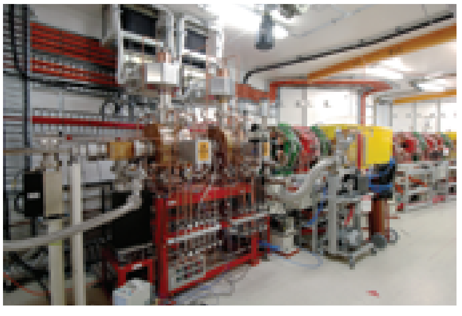


Figure 7.24
The booster to storage ring transfer line in the Australian Synchrotron

Recall also that the radius of the path of the electrons is given by $r = \frac{mv}{eB_\perp}$. As the RF booster gives the electrons more energy, their mass becomes greater. This would mean that the radius, *r*, becomes larger—and the electrons would hit the wall of the evacuated tube. To prevent this happening, the strength of the field ($B_\perp$) is increased at the same rate as the energy (and hence mass) increases. This all happens very quickly—as in 1 s the electrons make over one million revolutions of the storage ring!

The storage ring

The booster ring channels the electrons into the *storage ring*, a doughnut-shaped tube. In the Australian Synchrotron this ring has a radius of 34.3 m and a circumference of 216 m. Around this ring are 14 bending magnets each 1.7 m long and with a field strength of 1.3 T. These keep the electrons in the circular path. They are separated by 14 straight sections in which focusing magnets keep the electrons confined to a flat beam less than half a millimetre wide and only two-hundredths of a millimetre high.

A simple calculation based on $r = \frac{mv}{eB_\perp}$, using the normal mass of the electron and $v = c$, gives us a radius of curvature of only a little more than a millimetre. However, as we saw above, the mass of the 3 GeV electrons is almost 6000 times the rest mass and so the actual curvature is more like 6 m. Remember, however, that the bending magnets are only 1.7 m long and make up only a small proportion of the total 216 m storage ring, which is why the actual radius of the whole ring is much larger.

You may recall that James Maxwell made the revolutionary discovery that light travels as perpendicular, oscillating electric and magnetic fields. He predicted that an oscillating electric charge would produce an oscillating magnetic field. We are now familiar with the term 'electromagnetic radiation' to describe a band of electromagnetic waves ranging from radio waves, with long wavelengths, to gamma rays of very short wavelengths and high energy. In 1887, Heinrich Hertz demonstrated that accelerating charged particles in a primary loop could produce an electromagnetic wave that travelled to a second loop. Radio use relies upon oscillating electrons in an aerial producing electromagnetic radiation in the form of radio waves.

SYNCHROTRON RADIATION, or synchrotron light, is the term given to a range of electromagnetic radiation of wavelengths from approximately 10^{-3} to 10^{-10} m. This electromagnetic radiation is produced by charged particles such as electrons or protons as they travel at speeds close to that of light in a curved path. The beam of synchrotron radiation produced falls in the shape of a cone ahead of the travelling charged particles.

In the storage ring of the synchrotron, electrons orbit for hours at a time at speeds near that of light. A series of magnets makes them bend in arcs as they travel through the ring. It is as the electrons change direction (i.e. accelerate), that they emit synchrotron radiation.

Several different types of magnets are used to direct the beam of charged particles. Bending magnets called dipole magnets guide the particles through small arcs that combine to produce 360° of bending around the ring. Other magnets called quadrupole and sextupole magnets refocus the beam to prevent it diverging and keep the particles on stable orbits. Steering and corrector magnets correct and fine tune the orbit to one-millionth of a metre! Under ideal conditions, with focusing elements incorporated, we can consider the dipole magnets as generating a uniform, vertical field which is kept constant in the storage ring.

The specific arrangement and strength of the magnets dictates the brightness, polarisation, energy distribution and coherence of the synchrotron light produced. This arrangement is called the lattice.

The force on the electron, $\boldsymbol{F}_{\mathrm{B}}$, in the storage ring acts in the horizontal plane, according to $\boldsymbol{F}_{\mathrm{B}} = q\boldsymbol{v}\boldsymbol{B}_{\perp}$, where $\boldsymbol{B}_{\perp}$ is the perpendicular magnetic flux density (T), q is the charge on the electron (C) and $\boldsymbol{v}$ is the velocity of electron (m s^{-1}).

As the electron beam circulates, it radiates synchrotron light and immediately begins to move to an orbit of smaller radius because it is therefore losing energy. During this time, the electrons pass through *RF cavities*, like those in the booster ring, to replenish the energy lost as they radiate electromagnetic radiation.

The RF cavities have electromagnetic fields oscillating at radio frequencies produced by amplifiers located next to the storage ring. These fields oscillate from positive to negative values extremely quickly, up to 500 million times per second (500 MHz). Electrons are accelerated by the electric fields that are induced in the cavities by the RF radiation. The oscillations of the RF radiation are synchronised with the orbital period of the electrons so that they arrive during the positive half of the electric field's cycle.

This is where the term 'synchrotron' originates. This process ensures that the electrons stay at a constant energy and remain stored in the ring.

Physics file

The Australian Synchrotron consists of 14 cells. Each cell contains two dipole magnets, six quadrupoles and seven sextupoles, to assist in the steering and focusing of the electron beam.

The characteristics of the RF system and parameters of the lattice arrangement determine the length, duration and spacing of the bunches of electrons travelling around the storage ring.

Despite the RF cavities, the beam is still not perfectly stable. All synchrotron beams will gradually reduce in intensity with time. Some electrons are lost in collisions between electrons and gas molecules in the near vacuum of the ring. To minimise these losses, the vacuum chamber must be kept at a pressure of about one-thousandth of one-billionth of normal atmospheric pressure, or less than 10^{-7} Pa. Under these conditions, the beam typically loses half of its intensity over a 5–50 h period. New electrons are injected into the beam at 4–24 h intervals to replace those lost through collisions and energy losses.

The unused high-energy X-rays given off by the storage ring are continually absorbed by radiation shielding. The shield wall surrounding the storage ring is usually made of lead and concrete. This tunnel completely encloses the storage ring, except for the *beamlines* through which radiation is guided. This design feature is critical for employee safety during synchrotron operation.

Beamlines

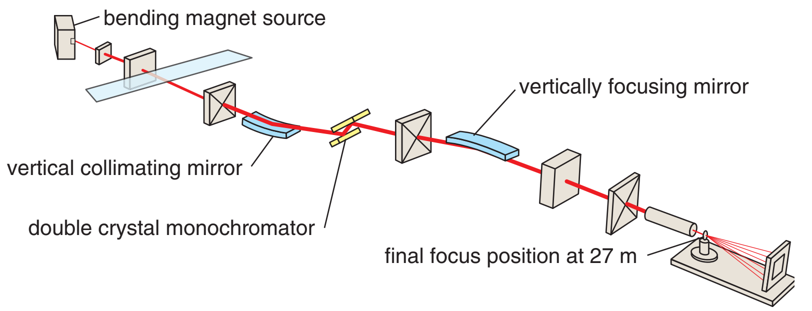


Figure 7.25

The arrangement of mirrors and crystal monochromator in the protein and microcrystal and small molecule X-ray diffraction beamline in the Australian Synchrotron.

A *beamline* is the path that synchrotron light travels from the storage ring, where it is produced, to its target experimental work. The point at which the beamline meets the storage ring is called the front end.

A beamline is typically a stainless steel tube, 15–35 m in length and around 4 cm in diameter. The dimensions depend greatly on the technique being performed on the beamline and the application of that technique. A typical beamline consists of an optics room, an experiment room and a control room.

Inside the optics room, synchrotron light is modified according to the needs of its experimental use. Sometimes scientists will wish to use only a specific range of wavelengths of synchrotron light for their experiments, rather than all of the light produced. A device called a *monochromator*, either a crystal or a grating, is used as a wavelength selector. As a beam hits this device, particular wavelengths are diffracted at different angles. By rotating the monochromator, a specific light frequency can be selected from the broad band of frequencies available in the incident beam. Synchrotron light may also be collimated using slits, refocused using mirrors, or lessened in intensity by using attenuators within the optics room as it is prepared for its role in an experiment.

Thin beams of synchrotron light are directed onto a specific target or sample being examined within an experiment room. Scientists may control their experiments from an external control room, in which they are protected from the intense electromagnetic radiation being used in the experiments. All synchrotrons have a number of beamlines, each directing the synchrotron light to an experimental station.

There are a number of ways that scientists can utilise synchrotron light in an experiment. The light transmitted through the sample can be analysed, as can that reflected off the front of it. Fluorescence processes can be investigated, as can the nature of electrons given off by secondary processes within the sample. A very powerful benefit of synchrotron light sources is that virtually any radiation wavelength can be used in experiments. Computers analyse the interactions of samples with synchrotron light so that results can be displayed using graphs or even three-dimensional images.

Table 7.2 on page 356 displays the initial 13 beamlines established at the Australian Synchrotron and an overview of some of the research fields that will benefit from the work on each of these.

Generations of synchrotron sources

Chances are that you share some characteristics with your parents and grandparents but you are still quite unique. In the same way, we can talk about different generations of synchrotron sources.

First-generation synchrotron sources were high-energy accelerators, built for research into the fundamental particles of matter. These machines accelerated charged particles around a circular orbit, until eventually they flew off at a tangent to their path, to collide with a target material. Nuclear scientists examined the fragments resulting from collisions to increase their knowledge of the structure of the atomic nucleus. Scientists realised that they were limited by the amount of energy they could give the particles, because as they moved through their orbit, the particles released unexpected electromagnetic radiation. The more energy they received, the more synchrotron light was produced. Because synchrotron light robbed the particles of energy, it was seen as an unwanted by-product of the process. It was not until the 1970s that the broad applications of synchrotron light began to be realised. Particle accelerators, originally designed to minimise synchrotron radiation, were now being created specifically to produce it.

These second-generation facilities were then constructed and became known as dedicated sources. The Synchrotron Radiation Source (SRS) at Daresbury in north-west England is an example of a second-generation source that is still in operation. Such sources were designed specifically to provide synchrotron X-rays for research. Typically, they consist of a large number of beamlines and experimental stations branching off the storage ring.

Over the last 20 years, a third generation of sources has developed. The electrons in the storage rings of these devices have energies ranging from 1 to 8 GeV. The diameter of the storage ring needs to be greater to attain these electron energy levels. Huge machines were built, including the European Synchrotron Radiation Facility (ESRF) in France, the Advanced Light Source (ALS) in California, USA, and SPring-8, the Super Photon Ring in Japan. Some synchrotrons currently being built are more compact, with circumferences of 100–200 m and electron energies around 3 GeV. The Australian Synchrotron fits into this category.

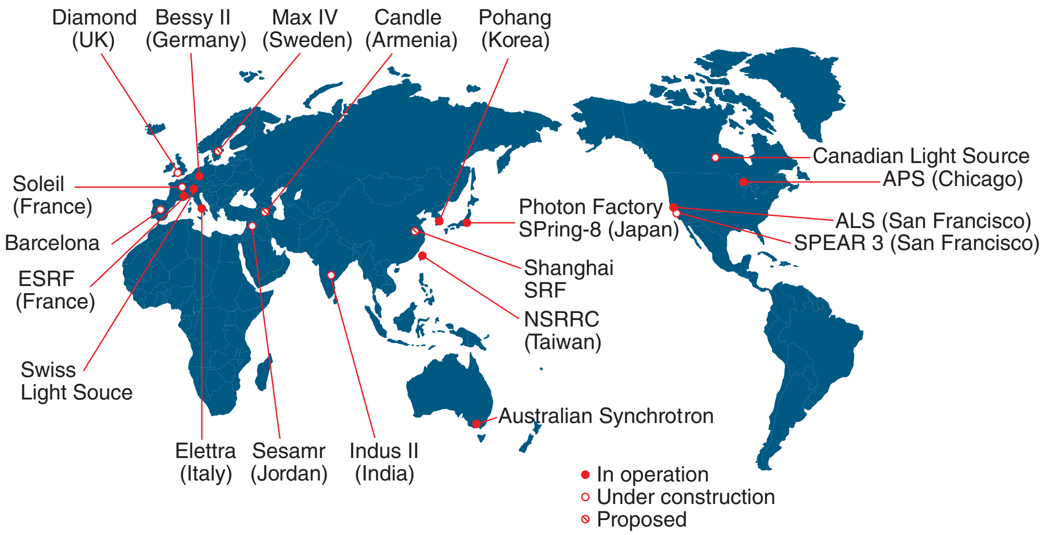


Figure 7.26

This map identifies third-generation synchrotron facilities with energies greater than 1.5 GeV currently proposed and in operation around the world.

Physics file

In 2002, the International Machine Advisory Committee (IMAC) considered an enhancement to the original design of the Australian Synchrotron that was completed by Professor John Boldeman. This enhanced design was approved by an international panel of experts. The Australian Synchrotron is now built to be twice as bright as originally planned. It is a 3 GeV, third-generation synchrotron, with a storage ring with a circumference of 216 m. The construction of the medium-energy 3 GeV capacity was viewed as the best value for money synchrotron design, with sufficient energy for a wide range of experiments and sufficient flexibility to support further developments. Twelve straight sections in the ring provide space for the insertion devices that increase the brilliance of the light. This light can be delivered to a capacity of around 30 beamlines. Although construction of the facility cost around $206m, an independent economic survey has suggested it will stimulate $21.7b in industrial development over its 25-year lifetime, and create 3500 new jobs. It will become Australia's premier research facility and attract hundreds of researchers each year, many from other Pacific Rim countries such as New Zealand, South Africa, Malaysia and Singapore. The synchrotron benefits Australian pharmaceutical development, the mining and mineral exploration industries and the manufacture of microstructure products, among many others (see Table 7.2) on page 356.

Third-generation sources are about 10 000 times brighter than second-generation sources. They derive this increase in intensity from insertion devices, called undulators and wigglers, which are placed in straight sections of the storage ring. Improvements in synchrotron components and technology have enabled much better performance from smaller synchrotrons than previously possible. In 2003, there were 52 synchrotron facilities worldwide, with most scientifically developed countries enjoying access to a local facility.

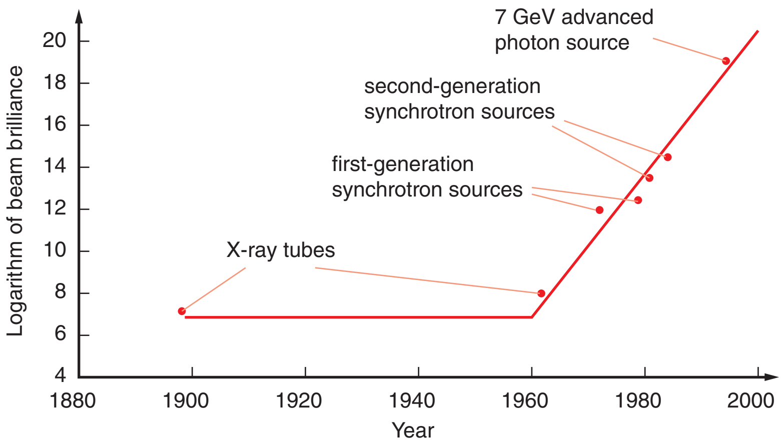


Figure 7.27

The brightness of subsequent generations of X-ray sources has greatly increased over the past 50 years.

Insertion devices

Synchrotron sources provide much brighter X-ray beams than traditional X-ray tubes. The brilliance of the beam is a measure of its intensity. The intense brightness of synchrotron light produced in a third-generation source is largely due to the effect of many small magnets called *insertion devices*. The two main types of insertion devices are the *wiggler* and the *undulator*. These are inserted within straight sections of the storage ring, between existing dipole (bending) and corrector magnets. Twelve of the total of 14 straight sections of the Australian Synchrotron are suitable to fit insertion devices.

Recall that the basic dipole magnet bends the path of the electrons in arcs. At the deflection of the electron beam, a cone of synchrotron light is produced, as shown in Figure 7.28a.

A wiggler consists of two rows of small alternating magnetic poles. These force the electron beam into a series of deflections. A cone of synchrotron light is emitted at each peak in the deflection. This light reinforces to increase the intensity of synchrotron light produced, much like a line of torches shining in the same direction. The resulting radiation is increased in intensity and brightness by a factor approximately equal to two times the number of magnetic poles. That is, a wiggler consisting of six poles will increase the brightness of synchrotron light produced 12 times. The reinforced synchrotron light emerges from the wiggler as a broad band of incoherent radiation, as can be seen in Figure 7.28b.

An undulator consists of less powerful magnets than that of the wiggler. These produce gentler deflections of the electron beam. The emitted synchrotron light overlaps to produce a beam that is collimated to a narrow width, as can be seen in Figure 7.28c. Undulators are partially monochromatic (or quasi-monochromatic) sources. This insertion device results in interference effects that produce a spectrum of synchrotron light that is enhanced at specific wavelengths. These wavelengths are determined by the spacing between poles of the undulator. In this respect, apart from greater brightness, undulator output is different from the continuous spectrum of the bending magnet or wiggler. For the fundamental wavelengths produced, the brightness can be one million times that produced by a bending magnet.

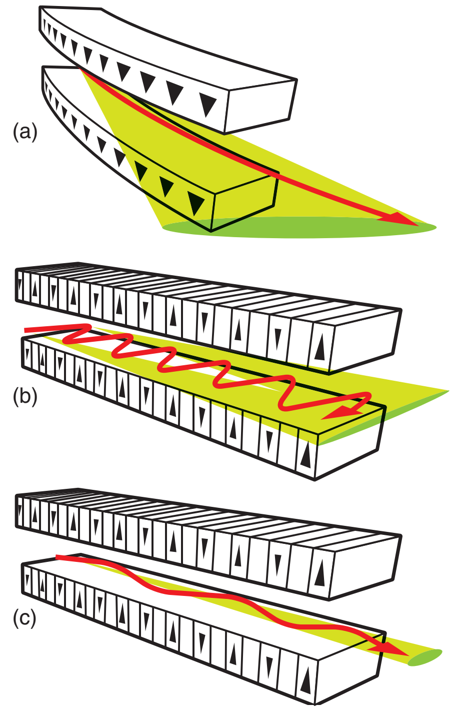


Figure 7.28

These diagrams show the bending of the path of the electrons and the resultant band of synchrotron light produced using (a) a dipole magnet, (b) a wiggler and (c) an undulator.

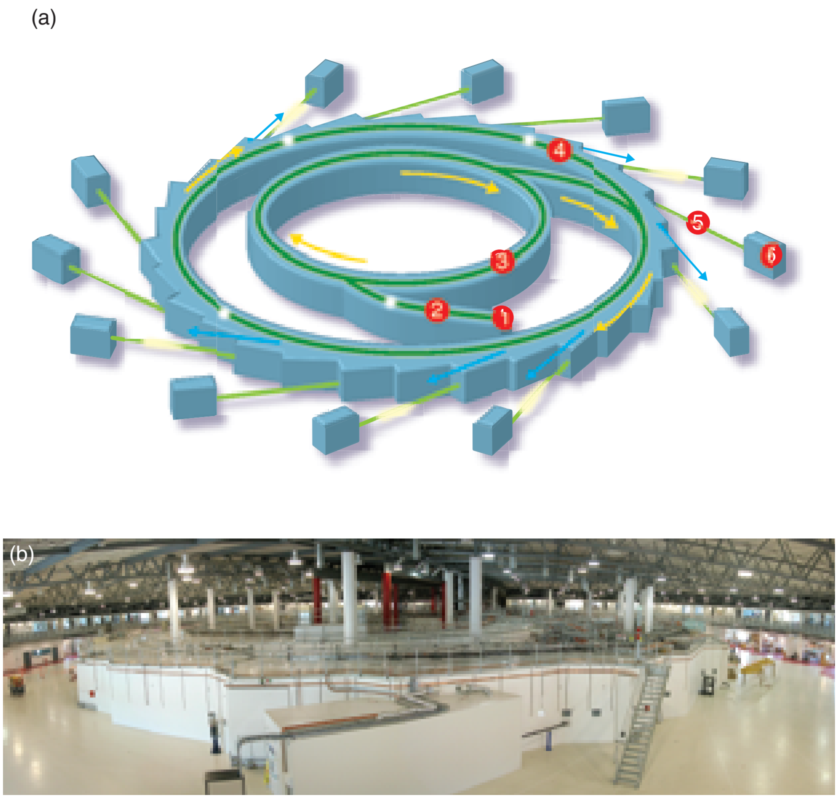


Figure 7.30

(a) The design of the Australian Synchrotron: (1) electron gun, (2) linac, (3) booster ring, (4) storage ring, (5) beamline, (6) experimental station. (b) View of the inside of the Australian Synchrotron, taken from the mezzanine.

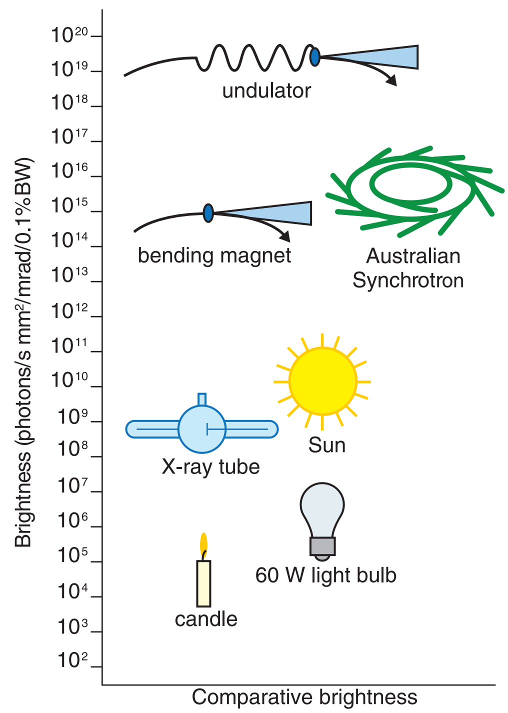


Figure 7.29

Compare the brightness of electromagnetic radiation produced by more traditional sources with what is possible using insertion devices within a synchrotron facility.

table 7.2 Applications of the Australian Synchrotron

Research fields	Protein crystallography	Protein micro-crystals and small molecules	Powder diffraction	Small and wide angle scattering	X-ray absorption spectroscopy	Soft X-ray spectroscopy	Vacuum UV	Vibrational and optical spectroscopy	Microfocus spectroscopies	Imaging and medical therapy	General purpose microprobe	Circular dichroism	Lithography
Life sciences													
Biological research and drug design	•	•	•	•	•			•	•	•	•	•	
Biotechnology and bio-sensors	•	•			•	•		•	•				•
Biomedical and medical imaging								•	•	•			
Medical therapy										•			
Plants and crops	•		•		•		•			•		•	
Physical sciences													
Sustainable environment		•	•		•			•	•				
Forensics			•	•	•	•		•	•	•			
Advanced materials													
– functional polymers		•		•		•		•		•		•	
– ceramics			•	•	•	•		•	•	•	•		
– nanomaterials and composites		•	•	•		•	•	•	•	•	•		•
– metals and alloys			•			•	•		•	•	•		
– micro-electronic and magnetic materials					•	•	•	•		•			•
– biomaterials		•	•		•	•	•	•		•	•	•	
Engineering			•							•			•
Mineral exploration and beneficiation			•		•	•			•	•	•		
Earth sciences		•	•	•	•	•		•	•		•		
Oil and gas production and distribution		•		•	•			•	•	•	•		
Agricultural technology	•		•	•	•	•				•		•	•
Food technology	•	•	•	•				•				•	•
Chemical reactions and catalysts		•	•	•	•	•	•	•	•		•		•

7.3 SUMMARY Synchrotrons

- A synchrotron is a doughnut-shaped (called a torus) particle accelerator designed to circulate electrons around a closed path at speeds very close to that of light.
- Electrons are emitted from an electron gun in pulses. They are accelerated in bunches through the linac by powerful bursts of RF (radio-frequency) radiation.
- Electrons then travel around a booster ring, being accelerated further as they pass through RF cavities until reaching an energy of 3 GeV. The magnetic field of the bending magnets is periodically increased as the velocity of electrons increases within the booster ring.
- Electrons are channelled into the storage ring. Synchrotron radiation is produced as the particles travel through the strong magnetic fields of the dipole magnets or insertion devices.
- Electrons replace energy lost due to the production of synchrotron light as they pass through RF cavities in the storage ring.
- Synchrotron light leaves the ring through front ends, passing down beamlines to a number of independent experimental stations. The beamline consists of an optics room, an experiment room and a control room from which the scientists monitor their experiment.
- Third-generation synchrotron sources are some 10 000 times brighter than second-generation sources owing to the use of insertion devices, called undulators and wigglers, placed in straight sections of the storage ring.

7.3 Questions

1 a Carefully describe the function of these major parts of the synchrotron:
 i linac
 ii circular booster ring
 iii storage ring
 iv beamline.
 b i In which section is the strength of the magnetic field increased as the velocity of electrons increases? Why is this necessary?
 ii In which section do the electrons move from a heated filament?
 iii In which section(s) are electrons accelerated by RF radiation?
 iv In which section would insertion devices be located?

2 a Define what is meant by the lattice of the storage ring.
 b Why is this important to synchrotron operation?

3 a Why are the RF cavities said to give particles an ‘energy boost’?
 b What would happen if a synchrotron was built without such RF cavities?
 c Explain how the presence of the RF cavities enables the path of charged particles to maintain an orbit of constant radius in the synchrotron, whereas particles in a cyclotron follow a spiral path.

4 The pressure inside the vacuum tube of the storage ring must be less than 10^{-7} Pa. Explain why this is important to its operation.

5 The principal function of a wiggler is to:
 A collimate the beam of synchrotron radiation produced
 B increase the brightness of synchrotron radiation produced
 C shift the wavelength of synchrotron radiation produced
 D deflect the orbit of synchrotron radiation produced.

6 Describe the design of an undulator and explain how its output of synchrotron radiation is different from that produced by a bending magnet or a wiggler.

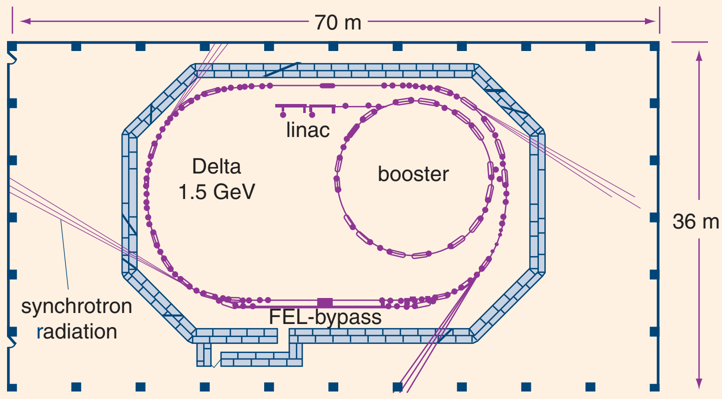
7 Read the Physics file on the Australian Synchrotron (page 354) to answer the following questions.
 a What is the energy of the electrons travelling in the Australian Synchrotron?
 b Study the suite of beamlines at the Australian Synchrotron shown in Table 7.2. Name three industries and describe how they may benefit from the development of the Australian Synchrotron.
 c Use the Victorian Government website to find three examples of recent benefits to industry that have been made by research teams using the Australian Synchrotron.

8 a Calculate the predicted velocity with which an electron will exit an electron gun with accelerating potential 120 kV, ignoring the effects of relativity.
 b Do you think the electrons will reach this speed? Explain.

9 Consider an electron, travelling at 99.999 998 55% of the speed of light through the storage ring of the Australian Synchrotron facility.
 a Given that the electron is travelling perpendicular to a uniform magnetic field of strength 1.5 T, calculate the expected radius of its bending path through this section, ignoring the effects of relativity.

b Do you think this is a realistic calculation of the true path radius of electrons in the Australian Synchrotron?

10 Study this diagram of the Delta synchrotron facility in Germany. Compare it with the Australian Synchrotron. List two similarities and two differences.



7.4 Mass spectrometry

> **Physics file**
>
> Particles of different masses follow different pathways in a mass spectrometer. Consequently a mass spectrometer can be used as an 'electromagnetic separator' to separate atoms of different masses in a mixture. This is how isotopes can be identified—and separated. South Australian born physicist Mark Oliphant recognised in the late 1930s that uranium isotopes could be separated, and in 1940, Otto Frisch and Rudolf Peierls calculated that a fission bomb of uranium-235 was a possibility. In 1943, Oliphant moved to work on the Manhattan Project, joining a research group on electromagnetic separation of uranium-235 from uranium-238 for the atomic bomb. Prior to this Mark Oliphant had discovered tritium and helium-3. He also developed vastly improved radar for the British—improvements that probably saved Britain from invasion by Germany during the famous 'Battle of Britain'. Mark Oliphant was the foundation director of the Division of Physical Sciences at the Australian National University and was also Governor of South Australia from 1972 to 1977.

How might it be possible to measure the mass of an atom? It is possible by using some really quite simple techniques to control the movement of things as small as atoms.

The mass spectrometer measures the mass of charged particles by measuring the path they follow in a magnetic field. Figure 7.31 shows the arrangement for a Bainbridge mass spectrometer. As already detailed earlier, moving charges in a magnetic field experience a force at right angles to their velocity.

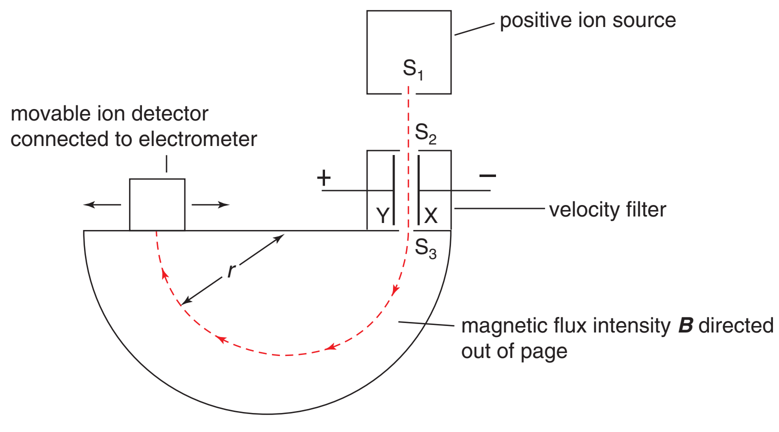


Figure 7.31
The Bainbridge spectrometer. Ions of uniform velocity enter a magnetic field at 90° to the field and continue in a circular path.

The force on a charged particle in a magnetic field depends on the size of the charge, the intensity of the field and the speed of the particle:

$$\boldsymbol{F} = qv\boldsymbol{B}_{\perp}$$

where $\boldsymbol{F}$ is the force (N), q is the charge (C), v is the speed (m s^{-1}) and $\boldsymbol{B}_{\perp}$ is the magnetic flux density at right angles to the motion of the charged particle (T).

Since the force is always at 90° to the velocity, the force is centripetal and will cause circular motion. The centripetal force is the magnetic force.

Mathematically the situation can be analysed as follows.

The ***CENTRIPETAL FORCE*** is a magnetic force:

$$F_c = F_B$$

$$\frac{mv^2}{r} = qv\boldsymbol{B}_{\perp}$$

$$m = \frac{rq\boldsymbol{B}_{\perp}}{v}$$

where r is the radius of the path (m), q is the charge on the particle (C), m is the mass (kg) and $\boldsymbol{B}_{\perp}$ is the perpendicular magnetic field of flux density (T).

All of the quantities on the right of the above equation can be measured. By simple substitution, the mass of the particle can be calculated.

The intensity of the magnetic field can be found very easily from the geometry of the coils creating the field and the size of the current in those coils. The formulas for these are not given here as they are not used in this course. The charge on the particles will be simple multiples of the charge on one electron. This value of 1.6×10^{-19} C is well established and has been accurately measured many times. The velocity of the charged particles can be controlled.

The particles are accelerated by a voltage before they enter the magnetic field. The kinetic energy of a particle that has been accelerated by a voltage just depends on the size of the charge and the size of the voltage:

$$E_k = q\Delta V$$

where E_k is the maximum kinetic energy of the accelerated particles (J), q is the charge of the particles (C) and ΔV is the accelerating voltage (V).

The particles accelerated by a voltage will have a range of velocities. Since the paths followed by the particles depend on their speed as well as their mass it is necessary to filter the beam of particles so that only those of one particular velocity enter the magnetic field of the spectrometer. This is done with a pair of 'crossed fields', one electric and one magnetic. Figure 7.32 shows the arrangement of magnetic and electric field that will allow only charges of one speed to travel through in a straight line.

The force on the charged particles due to the electric field is just qE where E is the intensity of the electric field in volts per metre. The force due to the magnetic field is $qv\boldsymbol{B}_{\perp}$.

If the force from each field is to be equal and opposite to the other then:

$$\boldsymbol{F}_E = \boldsymbol{F}_B$$

$$q\boldsymbol{E} = qv\boldsymbol{B}_{\perp}$$

$$v = \frac{\boldsymbol{E}}{\boldsymbol{B}}$$

where v is the speed of the particles which travel straight through the crossed fields (m s^{-1}), $\boldsymbol{E}$ is the intensity of the electric field (V m^{-1}) and $\boldsymbol{B}_{\perp}$ is the flux density of the perpendicular magnetic field in tesla (T).

As shown in Figure 7.32, particles travelling faster or slower than this speed will experience greater or lower force from the magnetic field and will be deflected by the combined effect away from the straight path into the spectrometer.

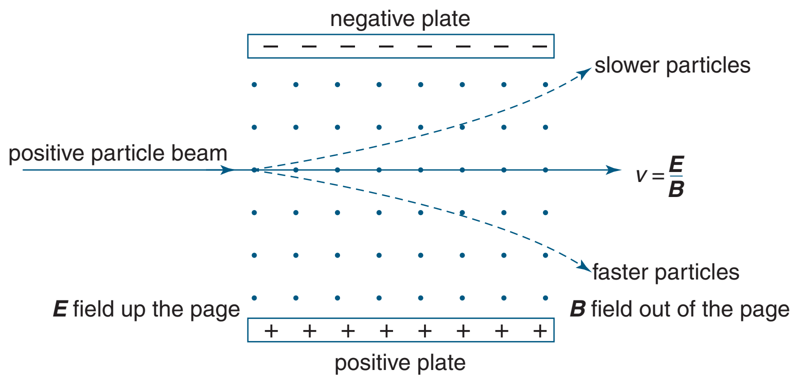


Figure 7.32

The crossed fields arrangement for a velocity selector. Only charged particles of one particular velocity will pass through the fields in a straight line. Faster or slower particles will travel in different paths because the size of the magnetic force depends on the speed of the particles, whereas the electric force is independent of the speed of the particles.

The radius of the path taken by the particles is easily measured once the particles have been detected at the end of their semicircular pathway. There are various ways of detecting the ions at the end of their curved path. Historically, this was done with photographic film, but now it is done with electronic ion or charge detectors.

Physics in action — Too small to measure?

Mass spectrometers are a large, diverse family of instruments that are used in many applications. Possibly no other type of complex instrument has been as important for so many areas of science and technology in the last 100 years. The many designs of mass spectrometers have an impressive range of applications.

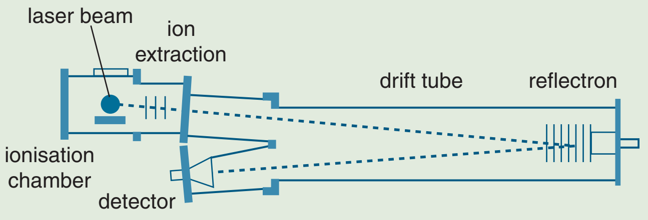


Figure 7.33

A time-of-flight mass spectrometer showing removal of ions from a solid sample with a laser. The reflectron is a series of rings or grids that act as an ion mirror. This mirror compensates for the spread in kinetic energies of the ions as they enter the drift region and improves the resolution of the instrument. The output of an ion detector is displayed on an oscilloscope as a function of time to produce the mass spectrum.

A mass spectrometer is any instrument that produces a mass spectrum—a record of the distribution of particles of different masses that can be found when a sample of the particles is ionised. After mass spectrometers were first made early in the 20th century, they proliferated both in type and in number. Fundamentally, mass spectrometers use the difference in mass-to-charge ratio (m/e) of ionised atoms or molecules to separate them from each other. Mass spectrometry is useful for the quantification of atoms or molecules and also for determining chemical and structural information about molecules. When molecules are broken up and ionised, the distinctive fragments provide structural information that can identify the original molecule.

In general, mass spectrometers create gaseous ions before separating them in space or time on the basis of their m/e ratio and then measuring the quantity of ions of each m/e ratio. Consequently, mass spectrometers always consist of an ion source, a mass-selective analyser and an ion detector.

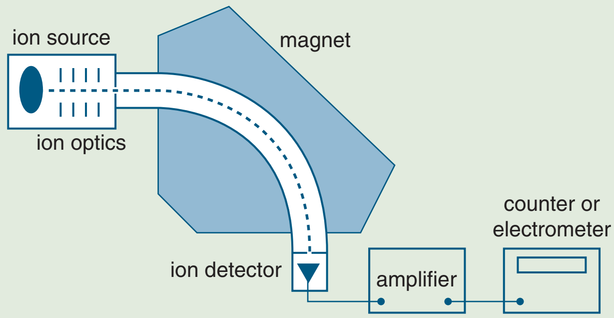


Figure 7.34

A curved pathway taken by ions in a magnetic field. The relationship between magnetic field intensity and the radius of curvature of the path enables identification of the ion masses.

Two common methods of ion separation are magnetic sector and time of flight. Magnetic-sector separation is described in the main body of the text.

A time-of-flight mass spectrometer separates ions according to the different times taken for them to travel through the mass spectrometer. An electric field accelerates all ions into the chamber of the spectrometer. The kinetic energy of an ion is given by

$$E_k = \frac{1}{2}mv^2 = q\Delta V$$

where m is the mass of the ion (kg), v is its velocity (m s^{-1}), q is the charge of the ion (C) and ΔV is the accelerating potential difference (V).

It can be seen that lighter ions will have a higher velocity than heavier ions and reach the detector at the end of the chamber sooner.

A magnetic-sector mass spectrometer uses the relationship between magnetic field and the ions' radius of pathway. The ion-source chamber of a mass spectrometer extracts and accelerates ions to a kinetic energy given by $q\Delta V$. The ions enter a curved evacuated tube between the poles of a magnet and are deflected by the perpendicular magnetic field $\boldsymbol{B}_\perp$. The ion detector at the other end of the tube will detect only ions of a particular m/e ratio. Changing either $\boldsymbol{B}_\perp$ or ΔV can vary the m/e of the ions that reach the detector. Magnetic-sector mass spectrometers typically vary the magnetic field through a range of values and record the field values at which ions are detected. This reveals the various masses of the ions.

By the early 1960s the measurement of isotopic masses and abundances allowed physicists and chemists to create new international standards for atomic mass. Capable of determining the mass of an atom to one part in a billion, mass spectrometers provide extremely precise measurements on large molecules and only need very small samples of material. In the early 1980s mass spectrometers could measure ions with molecular weights of over 10 000. By the 1990s the proven range had extended to several hundred thousand.

✓ Worked Example 7.4A

A beam of singly charged nitrogen molecules is fired into a mass spectrometer after being filtered through crossed fields of 4550 V m^{-1} and 0.120 T. What would be the radius of the semicircular path that they would follow in the spectrometer? The mass of nitrogen molecules is 4.68×10^{-26} kg.

Solution

First, find the velocity of the molecules:

$\boldsymbol{E} = 4550 \text{ V m}^{-1}$

$\boldsymbol{B}_\perp = 0.120 \text{ T}$

$$v = \frac{\boldsymbol{E}}{\boldsymbol{B}_\perp} = \frac{4550}{0.120} = 3.79 \times 10^4 \text{ m s}^{-1}$$

Now we can calculate the radius. Note that a singular charge is 1.60×10^{-19} C:

$m_N = 4.68 \times 10^{-26}$ kg

$\boldsymbol{B}_\perp = 0.120$ T

$v = 3.79 \times 10^4 \text{ m s}^{-1}$

$q = 1.60 \times 10^{-19}$ C

$$r = \frac{mv}{q\boldsymbol{B}_\perp} = \frac{(4.68 \times 10^{-26})(3.79 \times 10^4)}{(1.60 \times 10^{-19})(0.120)} = 9.24 \times 10^{-2} \text{ m}$$

Hence, the radius is equal to 9.24 cm.

Physics file

The first mass spectrometer was built in 1918 by William Aston. He was a student of J. J. Thomson, who was the first person to measure the mass-to-charge ratio of cathode rays, thus showing that they are negatively charged particles. Ions of the atoms being measured were produced in the glass bulb before a voltage accelerated them towards a magnetic field being produced by the coils visible at the left of the apparatus. The high voltages needed were produced by the induction coil at the bottom of the apparatus. The induction coil is a kind of transformer and works in exactly the same way as the 'coil' of a motor car ignition system.

Figure 7.35
F. W. Aston's mass spectrograph

✓ Worked Example 7.4B

In a mass spectrometer, germanium atoms have radii of curvature equal to 21.6 and 22.8 cm. The larger radius corresponds to an atomic mass of 76.0 u. What is the atomic mass of the other isotope?

Solution

The radius of the pathway is proportional to the mass of the particles. q, $\boldsymbol{B}$ and v are the same for each of the isotopes:

$$m = \frac{rq\boldsymbol{B}_{\perp}}{v}$$

$$\frac{m}{r} = \frac{q\boldsymbol{B}_{\perp}}{v}$$

This is constant for this problem so:

$$\frac{m_1}{r_1} = \frac{m_2}{r_2}$$

$$m_2 = \frac{r_2 m_1}{r_1}$$

$$m_2 = \frac{(76.0\text{u})(21.6)}{22.8}$$

$$m_2 = 72.0\text{ u}$$

Hence, 72 u is the atomic mass of the other isotope.

7.4 SUMMARY Mass spectrometry

- Moving charged particles follow a circular path in a magnetic field. The centripetal force is the magnetic force and is at 90° to the velocity of the particle.
- The radius of the path depends on the mass of the particle $m = \frac{rq\boldsymbol{B}_{\perp}}{v}$, where m is the mass of the particles, r is the radius of the path they follow, $\boldsymbol{B}_{\perp}$ is the magnetic flux density of the perpendicular magnetic field and v is the speed of the particles.
- Since the radius of the path can be measured easily with an ion detector and q, $\boldsymbol{B}_{\perp}$ and v can be deduced from other measurements, the mass of a particle can be calculated.
- The velocity of the particles entering a mass spectrometer can be controlled with crossed electric and magnetic fields. The velocity of the particles entering the spectrometer is E/B where E is the electric field intensity in V m^{-1} and B is the magnetic field intensity in tesla (T).

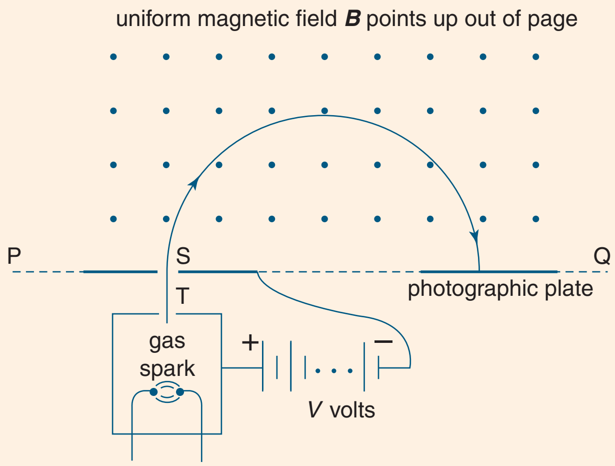
7.4 Questions

1 Protons with a speed of 1500 m s^{-1} enter the magnetic field of a mass spectrometer of size 0.600 T. What is the radius of their path?

2 What is the radius of the path of a proton which enters the magnetic field of a mass spectrometer of size 0.450 T, if they have been accelerated across a voltage of 2550 V?

3 A beam of singly positively charged oxygen atoms was accelerated by a potential difference of 1.10×10^4 V and enters the magnetic field of a mass spectrometer. If the field has a flux density of 7.20×10^{-2} T and the path radius is 0.850 m, calculate the mass of an oxygen atom.

The following information relates to questions 4–7.



The diagram represents, schematically, a Dempster mass spectrograph. A spark discharge in a gas produces ions of negligible kinetic energy. These drift through a hole T and are accelerated by an electric field, produced by a battery of *V* volts. The ions pass through the hole S, located in the plane PQ, and enter the uniform magnetic field. The ions then follow a semicircular path and are recorded on a photographic plate, also in the plane PQ. The whole apparatus is in a vacuum.

4 Are the ions in this case positive or negative?

5 Without altering the position of the photographic plate, what adjustments would have to be made to the apparatus in order to detect ions carrying an opposite charge?

6 What must be the voltage of the battery to give a proton a speed of 2.00×10^6 m s^{-1} as it passes through S?

7 If the radius of the proton's path is 0.555 m, what is the magnetic flux density in the region shown?

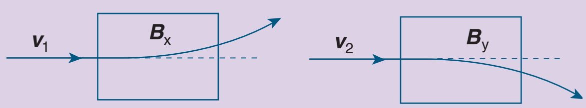
8 Would it be feasible to compare the radius of the paths of an electron and a proton under the same conditions in a mass spectrometer?

9 When a mass spectrometer is used to monitor air pollutants it is difficult to distinguish between molecules with very similar masses such as CO (28.0106 u) and N_2 (28.0134 u). Show that the radius of curvature of the mass spectrometer must be about 1.65 m if these two molecules are to be separated by 0.330 mm when they reach the detector.

Chapter 7 Review

Assume that:
charge on an electron, $e = 1.60 \times 10^{-19}$ C
mass of an electron, $m = 9.11 \times 10^{-31}$ kg
Planck's constant, $h = 6.63 \times 10^{-34}$ J s^{-1}

1 The following diagram shows two different electron beams being bent as they pass through two different regions of uniform magnetic field of equal magnitudes $\boldsymbol{B}_{\perp x}$ and $\boldsymbol{B}_{\perp y}$. The initial velocities of the electrons in the respective beams are $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

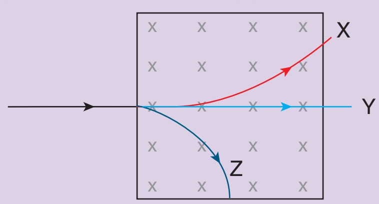


Which of the following is correct?

A $v_1 = v_2$	$\boldsymbol{B}_{\perp x}$ into the page	$\boldsymbol{B}_{\perp y}$ out of the page
B $v_1 = v_2$	$\boldsymbol{B}_{\perp y}$ into the page	$\boldsymbol{B}_{\perp x}$ out of the page
C $v_2 > v_1$	$\boldsymbol{B}_{\perp y}$ into the page	$\boldsymbol{B}_{\perp x}$ out of the page
D $v_2 < v_1$	$\boldsymbol{B}_{\perp y}$ out of the page	$\boldsymbol{B}_{\perp x}$ into the page

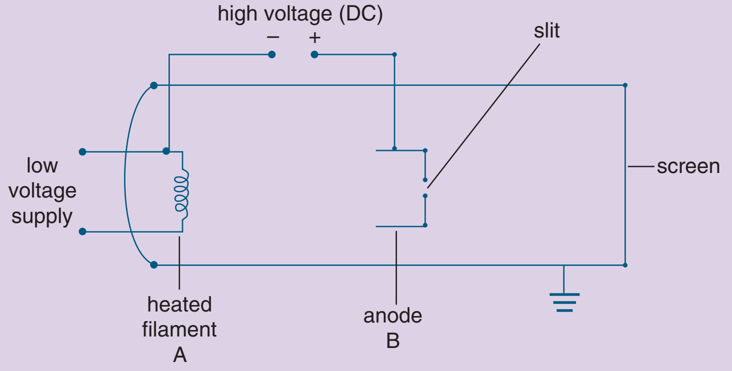
2 Justify your answer to Question 1.

3 The following diagram shows the paths taken by three different particles, X, Y and Z, as they pass through a uniform magnetic field.



Which of the following could be correct?
A X is a proton, Y is an electron, Z is a neutron.
B X is an electron, Y is a neutron, Z is a proton.
C X is a proton, Y is a neutron, Z is an electron.

4 Study the diagram (right) of a simple cathode ray tube. The source of electrons in this device is the:
A heated filament at A
B positive anode at B
C wires used in the circuit
D screen used in the circuit.



5 A particular electron gun accelerates an electron across a potential difference of 15.0 kV, a distance of 12.0 cm between a pair of charged plates. Calculate the magnitude of the force acting on the electron.

6 In an electron gun, electrons are accelerated from:
A a cathode towards a positively charged anode due to a magnetic field
B a cathode towards a positively charged anode due to an electric field
C an anode towards a negatively charged cathode due to a magnetic field
D an anode towards a positively charged cathode due to an electric field.

7 In an electron gun, an electron is accelerated by a potential difference of 28.0 kV. Calculate the velocity with which the electron will exit the assembly.

8 If the electron mentioned in Question 7 was accelerated a distance of 20.0 cm between a pair of charged parallel plates, then determine the size of the electric field.

9 Synchrotron light has which of the following characteristics:
A narrow spectral range, high intensity, highly polarised
B broad spectral range, high intensity, not polarised
C narrow spectral range, low intensity, not polarised
D broad spectral range, high intensity, highly polarised.

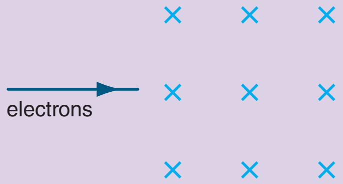
10 An electron travels through a magnetic field of strength 1.20 T with a velocity of 4.20×10^6 m s^{-1}. Calculate the radius of curvature of the path it will follow.

11 What is the energy in joules and electronvolts of a photon of wavelength 6.80×10^{-11} m?

12 The major difference between Thomson and Compton scattering is that X-rays involved in:
A Thomson scattering exit with no loss of kinetic energy to scattering atoms
B Compton scattering exit with no loss of kinetic energy to scattering atoms
C Compton scattering interact with bound electrons and cause the ejection of a photoelectron
D Thomson scattering interact with bound electrons and cause the ejection of a photoelectron.

13 X-ray photons of wavelength 9.80×10^{-10} m are incident on an aluminium sample material. Given that 4.20 eV is required to remove an electron, then calculate the maximum kinetic energy of ejected electrons.

14 This diagram shows a stream of electrons entering a magnetic field. Reproduce the diagram and show the subsequent path of the electrons through the magnetic field.



15 An electron beam travelling through a cathode ray tube is subjected to simultaneous electric and magnetic fields. The electrons emerge with no deflection. Given that the potential difference across the parallel plates X and Y is 3.00 kV, and that the applied magnetic field is of strength 1.60×10^{-3} T, calculate the distance between plates X and Y.

16 In an experiment similar to Thomson's for determining the charge-to-mass ratio, $\frac{q}{m}$, of cathode rays, electrons travel at right angles through a magnetic field of strength 1.50×10^{-4} T. Given that they travel in an arc of radius 6.00 cm and that $\frac{q}{m} = 1.76 \times 10^{11}$ C kg^{-1}, calculate the speed of the electrons.

17 Two ions, each singly charged, enter the magnetic field of a mass spectrograph at the same point with the same velocity. If their masses are in the ratio 5 : 3 and the path of the lighter particle has a radius of 155 mm, what is the distance apart of the points where the heavier ion enters and leaves the magnetic field?

18 Consider a spectrograph in which singly charged ions of helium are accelerated by a potential difference of 500.0 V and move in a circular path of radius 40.2 mm in a magnetic field of flux density 0.160 T. determine the mass of a helium ion.

19 A mass spectrometer is used to separate isotopes. A beam of singly positive charged oxygen atoms is accelerated by a potential difference and passed through a velocity selector. Atoms with 11.0 keV of kinetic energy are allowed into the main body of the spectrometer. The magnetic field inside the spectrometer has a flux density of 7.20×10^{-2} T and the path of radius is 0.850 m. Calculate the mass of the oxygen atoms.

Appendix

Formulas and constants

Forces and motion	
Mean velocity	$v_{av} = \frac{s}{t} = \frac{v+u}{2}$
Equations of motion	$a = \frac{\Delta v}{\Delta t}$; $s = ut + \frac{1}{2}at^2$; $v^2 = u^2 + 2as$; $v = u + at$
Force	$F = ma$
Weight force	$F = mg$
Momentum	$p = mv$
Change in momentum (impulse)	$F\Delta t = mv - mu$
Kinetic energy	$E_k = \frac{1}{2} mv^2$
Gravitational potential energy	$E_p = mgh$
Work done	$W = Fs = \Delta E$
Power	$P = \frac{W}{t} = \frac{\Delta E}{t} = Fv_{av}$
Centripetal acceleration	$a_c = \frac{v^2}{r}$
Centripetal force	$F_c = ma_c = \frac{mv^2}{r}$
Newton's Law of Universal Gravitation	$F = G\frac{m_1 m_2}{r^2}$
Gravitational field strength	$g = G\frac{M}{r^2}$
Moment of a force	$\tau = rF$

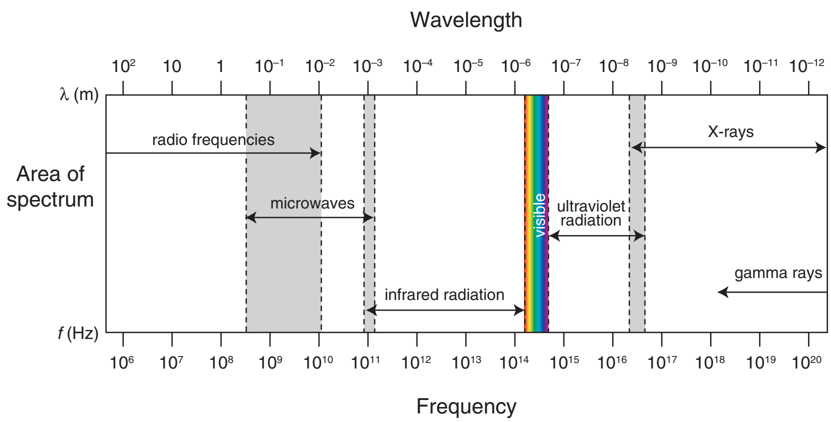
Electricity and magnetism	
Electric current	$I = \frac{q}{t}$
Electric field	$E = \frac{F}{q} = \frac{V}{d}$
Work and energy	$W = qV = VIt$
Ohm's law	$V = IR$
Resistances in series	$R_T = R_1 + R_2 + \ldots$
Resistances in parallel	$\frac{1}{R_T} = \frac{1}{R_1} + \frac{1}{R_2} + \ldots$
Power	$P = VI = I^2R = \frac{V^2}{R}$
Magnetic flux	$\Phi = BA$
Electromagnetic induction	$\text{emf} = -N\frac{\Delta\Phi}{\Delta t}$, $\text{emf} = lvB$
Magnetic force	$F = IlB$, $F = qvB$
Ideal transformer turns ratio	$\frac{V_s}{V_p} = \frac{N_s}{N_p}$

Particles and waves	
Energy of photon	$E = hf$
Energy transitions	$E_2 - E_1 = hf$
Wave period	$T = \frac{1}{f}$
Wave equation	$v_{wave} = f\lambda$
Internodal distance	$d = \frac{1}{2}\lambda$
Absolute refractive index	$n_x = \frac{c}{c_x}$
Snell's law	$n_1 \sin\theta_1 = n_2 \sin\theta_2$

Physical constants		
Speed of light in vacuum or air	c	$= 3.00 \times 10^{8}$ m s^{-1}
Electron charge	e	$= -1.60 \times 10^{-19}$ C
Mass of electron	m_e	$= 9.11 \times 10^{-31}$ kg
Planck's constant	h	$= 6.63 \times 10^{-34}$ J s
Universal gravitational constant	G	$= 6.67 \times 10^{-11}$ N m^2 kg^{-2}
Electron volt	1 eV	$= 1.60 \times 10^{-19}$ J
Mass of proton	m_p	$= 1.67 \times 10^{-27}$ kg
Mass of alpha particle	m_α	$= 6.65 \times 10^{-27}$ kg

Physical data		
Mean acceleration due to gravity on Earth	g	$= 9.80$ m s^{-2}
Mean acceleration due to gravity on the Moon	g_M	$= 1.62$ m s^{-2}
Mean radius of the Earth	R_E	$= 6.37 \times 10^{6}$ m
Mass of the Earth	M_E	$= 5.98 \times 10^{24}$ kg
Mean radius of the Sun	R_S	$= 6.96 \times 10^{8}$ m
Mass of the Sun	M_S	$= 1.99 \times 10^{30}$ kg
Mean radius of the Moon	R_M	$= 1.74 \times 10^{6}$ m
Mass of the Moon	M_M	$= 7.35 \times 10^{22}$ kg
Mean Earth–Moon distance	3.84×10^{8} m	
Mean Earth–Sun distance	1.50×10^{11} m	
Tonne	1 tonne	$= 10^{3}$ kg $= 10^{6}$ g

Electromagnetic spectrum

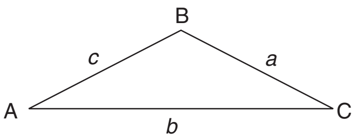


Note: 1. Shaded areas represent regions of overlap.
2. Gamma rays and X-rays occupy a common region.

Mathematical expressions

Given $ax^2 + bx + c = 0$, $x = \dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

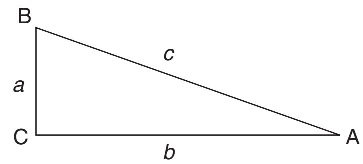
The following expressions apply to the triangle ABC as shown:



$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C}$$

$$b = \sqrt{a^2 + c^2 - 2ac\cos B}$$

The following expressions apply to the right-angled triangle ABC as shown:



$$\sin A = \frac{a}{c}$$

$$\cos A = \frac{b}{c}$$

$$\tan A = \frac{a}{b}$$

Some useful quantities

Time: 60 s = 1 minute
60 min = 1 hour
24 hours = 1 day
365.25 days = 1 year

Length: 1 angstrom unit = 10^{-10} m = 0.1 nm

Volume: 1 litre = 1 dm^3 = 10^{-3} m^3 = 10^3 cm^3

Mass: 1 tonne = 1 t = 10^3 kg
1 kilogram of water at 4°C occupies a volume of 1 litre

SI units and symbols

Fundamental SI units

Quantity	Symbol	SI unit	SI symbol
length	l	metre	m
mass	m	kilogram	kg
time	t	second	s
electric current	I	ampere	A
thermodynamic temperature	T	kelvin	K
luminous intensity	I	candela	cd
amount of substance	n	mole	mol

Some derived SI units, names and symbols

Quantity	Symbol	SI unit	SI symbol
velocity	$\boldsymbol{v}$	metres per second	m s^{-1} or m/s
acceleration	$\boldsymbol{a}$	metres per (second)^2	m s^{-2} or m/s^2
momentum	$\boldsymbol{p}$	newton second	N s
force	$\boldsymbol{F}$	newton	N
energy, work, heat	E, W, Q	joule	J
power	P	watt	W
electrical charge	q, Q	coulomb	C
electric potential, potential difference	V (ΔV)	volt	V
electric resistance	R	ohm	Ω
capacitance	C	farad	F
frequency	f	hertz	Hz
electric field strength	E	newtons per coulomb	N C^{-1} (= V m^{-1})
magnetic flux	ϕ	weber	Wb
magnetic field intensity	B	tesla	T (= Wb m ^{2} = N A^{-1} m^{-1})
gravitational field strength	g	newtons per kilogram	N kg^{-1} or N/kg
pressure	P	newtons per (metre)^2	Pa (= N m^{-2})
Young's modulus	E	newtons per (metre)^2	Pa (= N m^{-2})

Metric prefixes

Multiplying factor	Prefix	Symbol
10^{18}	exa	E
10^{15}	peta	P
10^{12}	tera	T
10^{9}	giga	G
10^{6}	mega	M
10^{3}	kilo	k
10^{-2}	centi	c
10^{-3}	milli	m
10^{-6}	micro	μ
10^{-9}	nano	n
10^{-12}	pico	p
10^{-15}	femto	f
10^{-18}	atto	a

Example: 1 GJ = 1×10^9 J

The periodic table of elements

Period \ Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Period 1									1 H 1.01									2 He 4.00
2	3 Li 6.94	4 Be 9.01											5 B 10.81	6 C 12.01	7 N 14.01	8 O 16.00	9 F 19.00	10 Ne 20.18
3	11 Na 22.99	12 Mg 24.31	Transition elements										13 Al 26.98	14 Si 28.09	15 P 30.97	16 S 32.06	17 Cl 35.45	18 Ar 39.95
4	19 K 39.10	20 Ca 40.08	21 Sc 44.96	22 Ti 47.90	23 V 50.94	24 Cr 52.00	25 Mn 54.94	26 Fe 55.85	27 Co 58.93	28 Ni 58.71	29 Cu 63.54	30 Zn 65.37	31 Ga 69.72	32 Ge 72.59	33 As 74.92	34 Se 78.96	35 Br 79.91	36 Kr 83.80
5	37 Rb 85.47	38 Sr 87.62	39 Y 88.91	40 Zr 91.22	41 Nb 92.91	42 Mo 95.94	43 Tc (99)	44 Ru 101.07	45 Rh 102.91	46 Pd 106.4	47 Ag 107.87	48 Cd 112.40	49 In 114.82	50 Sn 118.69	51 Sb 121.75	52 Te 127.60	53 I 126.90	54 Xe 131.30
6	55 Cs 132.91	56 Ba 137.34	57 La 138.91	72 Hf 178.49	73 Ta 180.95	74 W 183.85	75 Re 186.2	76 Os 190.2	77 Ir 192.2	78 Pt 195.09	79 Au 196.97	80 Hg 200.59	81 Tl 204.37	82 Pb 207.19	83 Bi 208.98	84 Po (210)	85 At (210)	86 Rn (222)
7	87 Fr (223)	88 Ra (226)	89 Ac (227)	104 Rf (261)	105 Db (262)	106 Sg (263)	107 Bh (262)	108 Hs (265)	109 Mt (266)	110 Ds (269)	111 Rg (272)	112 Uub (277)	113	114 Uuq (289)	115	116 Uuh (289)	117	118 Uuo (293)

Lanthanides

58 Ce 140.12	59 Pr 140.91	60 Nd 144.24	61 Pm (145)	62 Sm 150.35	63 Eu 151.96	64 Gd 157.25	65 Tb 158.92	66 Dy 162.50	67 Ho 164.93	68 Er 167.26	69 Tm 168.93	70 Yb 173.04	71 Lu 174.97

Actinides

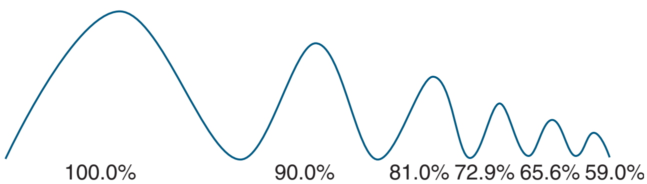
90 Th 232.04	91 Pa (231)	92 U 238.03	93 Np (237)	94 Pu (242)	95 Am (243)	96 Cm (247)	97 Bk (247)	98 Cf (249)	99 Es (254)	100 Fm (253)	101 Md (256)	102 No (254)	103 Lr (257)

Every isotope of these elements is radioactive

Answers

Fairground physics

E1 7.67×10^{-1} m s^{-1}
E2 A
E3 Outside column
E4 4.12×10^{2} N
E5 1.92×10^{2} N towards the centre
E6 4.47×10^{1} m s^{-1}
E7 7.51×10^{3} N down the slope
E8 1.02×10^{2} kg
E9 3.13×10^{1} m s^{-1}
E10 1.15×10^{5} J
E11 3.94×10^{1} m
E12 same
E13 1.84×10^{1} m
E14

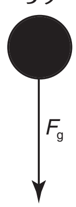


E15 1.84 m
E16 1.21×10^{1} m s^{-1}
E17 1.13×10^{4} N
E18 A 2.00g ride can be created if the accelerating force is vertical as opposed to horizontal, for example a 2.00g force can be experienced at the bottom of a curve.
E19 2.63×10^{1} N
E20 5.36×10^{-1} m s^{-2}
E21 **a** 5.92×10^{-1} m s^{-1} or 2.13 km h^{-1}
b -1.45×10^{-1} m s^{-1} or -5.22 km h^{-1}
E22 -3.26 m s^{-1} or -1.17×10^{1} km h^{-1}
E23 1.66 m s^{-1} or 5.98 km h^{-1}
E24 1.05 m s^{-1}

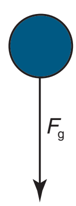
Chapter 1

1.1 Projectile motion

1 a 1.00 s **b** 20.0 m **c** -9.8 m s^{-2} **d** 18.4 m s^{-1} **e** 22.3 m s^{-1}
2 a 10.0 m s^{-1} **b** 4.43 m s^{-1} down
c 10.9 m s^{-1} 23.9° down from the horizontal **d** 0.452 s
e 4.52 m **f**



3 a 24.2 m s^{-1} **b** 24.2 m s^{-1} **c** 24.2 m s^{-1}
4 a 14.0 m s^{-1} up **b** 4.20 m s^{-1} up **c** 5.60 m s^{-1} down
5 a 1.43 s **b** 10.0 m **c** -9.80 m s^{-2}
6 a at its maximum height **b** 24.2 m s^{-1} **c** 1.43 s **d**



7 a 2.86 s **b** 19.8 m s^{-1} 45.0° down from the horizontal
c 69.3 m
8 C
9 a 6.53 m **b** 11.3 m s^{-1} up **c** 45.0° **d** 11.4 m s^{-1}
e 13.1 m 29.5° up from horizontal **f** 2.31 s **g** 26.1 m
10 22.9 m s^{-1} **b** 11.0 m **c** 3.57 s

1.2 Circular motion in a horizontal plane

1 a A
b She has continued to travel in a straight line, while the car has turned, so the right side of the cabin is actually accelerating towards her.
2 a 8.00 m s^{-1} **b** 8.00 m s^{-1} south **c** 6.96 m s^{-2} east
3 a 8.35×10^{3} N east
b The force that causes the centripetal force is the reaction of the sideways frictional force of the car's tyres on the road, that is the sideways force of friction of the road on the car's tyres.
4 a 8.00 m s^{-1} north **b** west
5 The car would probably skid off the road as the centripetal force required would increase to a value greater than the force of friction could provide.
6 a 2.67 m s^{-2} towards the centre
b The forces are unbalanced. According to Newton's first law, an unbalanced force will cause an object to change its motion, in this case the direction of the motion is changing, not the magnitude.
c 1.33×10^{2} N
d The sideways reaction force of the skate on the ice, which is the sideways force of the ice on the skate.
7 a 5.00×10^{-1} s **b** 10.1 m s^{-1}
c 1.26×10^{2} m s^{-2}
d 3.16×10^{2} N towards the centre
e The force causing the centripetal acceleration of the ball is the tension force of the cable on the ball.
f The ball would continue in a straight line that is tangential to the circular path at the point at which the wire breaks.
8 a 1.20 m
b The forces acting on Ella are gravity and the tension force of the rope on her.
c Ella's acceleration is towards the centre of rotation about the pole.
d 1.70×10^{2} N towards the centre **e** 2.61 m s^{-1}
9 a 4.86 kN **b** 22.5°
10 The driver will have to turn the car's tyres down the track to enable the horizontal component of the sideways

1.3 Circular motion in a vertical plane

1 a The acceleration is towards the centre of the circular path of the yo-yo.
b At the bottom of the circular path the tension in the string is greatest.
c At the top of the circular path the tension in the string is lowest.
d At the bottom of the circular path where the tension in the string is greatest.
2 3.83 m s^{-1}

3 a The force of gravity and the reaction force of the road on the car.
b 6.56×10^3 N upwards
c Yes it is possible, however it is her apparent weight she was 'feeling' not her mass, which doesn't change. The force that the seat applies to her is less as she goes over the hump, therefore she feels like she is lighter on the seat.
d 35.6 km h^{-1}
4 a 31.4 m s^{-1} **b** 19.9 m s^{-1} **c** 8.30×10^3 N downwards
5 12.1 m s^{-1}
6 1.96×10^2 N downwards
7 31.3 m s^{-1}
8 88.5 m s^{-1}
9 a 39.2 N **b** 1.18 N
10 The wire is more likely to break when the ball is moving through position X as the tension in the wire is three times the tension it had when it was stationary at point X.

Chapter review

1 a −0.50 m s^{-1} **b** 6.50 m s^{-1} upwards
2 a force increases until springs are fully stretched then decreases until Hannah leaves the trampoline; reaction force
b D
3 a 9.26 m s^{-2} **b** 2.78 m s^{-1} **c** 1.11 kN
4 a 0.639 s **b** 8.01 m s^{-1} **c** 9.77 m s^{-1} **d** 1.60 m
5 62.3°
6 4.65 m s^{-1}
7 D
8 8.40 m
9 a 4.00 m s^{-1} **b** 6.92 m s^{-1} **c** 0.707 s **d** 3.95 m **e** 1.60 s **f** 6.42 m
10 4.00 m s^{-1}
11 16.0 J
12 −9.80 m s^{-2}
13 C
14 2.13×10^{-3} N in the opposite direction of the motion
15 9.22×10^3; air resistance is insignificant when compared to the force due to gravity on the ball
16 a 2.80 kg m s^{-1} downwards **b** C **c** D **d** 4.80×10^3 N upwards
17 a 2.25×10^2 J **b** 2.25×10^2 J **c** 9.78×10^3 J
18 a 1.00×10^3 W **b** 9.78×10^2 W **c** 2.25×10^1 W
19 a 0 kg m s^{-1} **b** 2.00×10^2 kg m s^{-1} **c** -2.00×10^2 kg m s^{-1}
20 −1.00 m s^{-1}
21 −1.68 m s^{-1}
22 a -1.36×10^2 kg m s^{-1} **b** 1.36×10^2 kg m s^{-1}
23 1.00 m s^{-1} east
24 C
25 B
26 a A **b** D **c** C
27 12.6 s
28 a 5.00 m s^{-2} west **b** 7.55×10^3 N west
c 7.55×10^3 N east
29 C
30 a i 3.65×10^2 N upwards **ii** 6.15×10^2 N upwards **b** D

Chapter 2

2.1 Gravitational fields

1 a 5.34×10^{-12} N attraction **b** 1.64×10^5 N attraction
c 1.99×10^{20} N attraction **d** 3.61×10^{-47} N attraction
2 a $r_2 = 2 \times r_{Moon}$ **b** 10 N attraction
3 37.5
4 $r_2 = 10 \times r_E$
5 a 100 **b** $0.100R$
6 $0.990R$
7 $g_J = 24.8$ N kg^{-1}
8 a 296 N **b** 835 N **c** 1980 N
9 Saturn's mass is approximately 100 times that of Earth, while the radius of Saturn is only 10 times that of earth. When the radius of the planet is squared the factor of 10 becomes 102, which is enough to cancel out the factor of 100 by which the mass of Saturn is greater.
10 3.46×10^8 m

2.2 Satellite motion

1 D
2 As the satellite does not change its energy while orbiting around the Earth, It doesn't change its height above the surface of the Earth so its gravitational potential energy does not change, and its speed doesn't change so its kinetic energy doesn't change.
3 5.06×10^2 N
4 The source is the gravitational attraction of the Earth on the satellite.
5 a 0.0540 m s^{-2} **b** 4.38×10^3 m s^{-1} **c** 5.90 days
6 a 5.87 km s^{-1} **b** 2.67×10^{-2} m s^{-2}
7 6.67×10^{26} kg
8 a 1.54×10^9 m **b** 4.60×10^2 m s^{-1} **c** 1.37×10^{-4} m s^{-2}
9 a i 1.66 **ii** 7.58 **b** 15.6 days
10 6.98×10^6 m

2.3 Torque

1 a tap spindle, 0.04 m **b** axle of the wheel, 1 m
c tweezers, 0.07 m
d place in which the screwdriver contacts the edge of the tin, 0.2 m
2 a The line of application of the force is a larger perpendicular distance from the hinges (pivot point) when the force is applied to the handle than when it is applied to the centre of the door.
b Using a long crowbar with a small rock as a pivot a large force can be applied to the large rock if a small effort arm is used with a long effort arm, a ratio of load arm to effort arm of 1:10 will multiply the force you apply by ten times.
3 8.82×10^2 N m
4 2.50×10^2 N upwards
5 a 4.90×10^4 N m
b Cranes use counter-weights on the other side of the pivot point to the load to provide an opposing torque to balance the torque due to the load.
6 a 4.90 N m **b** 9.80 N m **c** 4.90 N m
7 The weights provide a large counter torque should the performer overbalance. Only a small movement of the pole is enough to balance the torque provided by the performer overbalancing.
8 The bench will not work successfully. The supports should be moved so that the centre of gravity is between the supports, or bolts could be used to attach the left hand support to the bench-top.
9 a The weight of the bag will produce a torque about a pivot point around the base of the spine, which will tend to rotate the torso to the right. To compensate for this the person must lean to the left, or by holding the left arm out from the body to move it farther from the pivot point.

b 40 N m

10 a 3.43×10^4 N

b As the perpendicular distance from the line of action of the load to the pivot point does not change, then the torque does not change. **c** 5.15×10^5 N m

2.4 Equilibrium

1 A, B and D

2 4.85×10^2 N

3 9.80×10^5 N

4 5.06×10^2 N

5 160 N

6 1.18×10^2 N m

7 a $\Sigma\tau_{acw} = 2F_t\, 10\sin 30°$ **b** $F_{t\,horiz} = F_t \cos 30°$, $F_{t\,vert} = F_t \sin 30°$
c 3.43×10^3 N

8 2.49 m

9 a 159 N m **b** 248 N m **c** 745 N m

10 a 578 N up **b** 1.65 m from the LHS

11 a 693 N and 400 N, respectively **b** 471 N

Chapter review

1 3.78×10^8 m from Saturn's centre of mass

2 D

3 a D **b** B **c** C **d** A **e** A

4 16

5 a 1.11×10^1 N kg^{-1} **b** C

6 27.4 days

7 a 3.39×10^3 m s^{-1} **b** 1.05×10^{-3} m s^{-2} towards Jupiter
c 236 days

8 a C
b satellite will always be positioned above the same location on Earth
c 4.23×10^7 m

9 a 2.43×10^8 m **b** 3.01×10^2 m s^{-1} **c** 3.73×10^{-4} m s^{-2}

10 a 2.75×10^7 J **b** 3.75×10^7 J **c** 1.94×10^3 m s^{-1} **d** 3.50 N kg^{-1}

11 a C **b** B **c** D

12 2.01×10^{30} kg

13 B

14 25.3 days

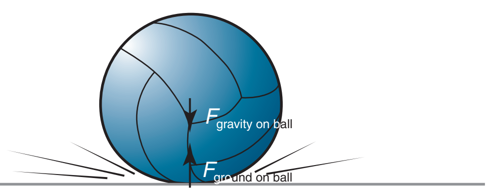
15 1.46×10^{22} kg

16 2.51

17 6.36×10^{-2}

18 39.4

19



20 a C **b** The forces must be on different objects and be equal in magnitude

21 278 N up

22 1.50×10^3 N m; hit the wall higher up

23 E

24 46.6 N m

25 B

26 B

27 22.9 N, 17.4 N

28 1.44×10^2 N, 1.60×10^2 N

29 a 1.88 m
b Reduces the torque acting on the crane, making it less likely it will topple over

30 a 250 N m **b** Y; the surface of the concrete at Y is under tension, while at X it is under compression

Chapter 3

3.1 Magnetic fields

1 B

2 C

3 a north **b** north-east **c** east

4 a south **b** north **c** zero

5 R

6 a south **b** south **c** south

7 a ***B***, into the page **b** 3***B***, into the page **c** zero

8 A

9 south

10 south

3.2 Force on current-carrying conductors

1 a 4.27×10^{-1} N upwards **b** 2.67×10^{-1} N downwards

2 918

3 B

4 a 2.00×10^{-4} N north **b** 1.00×10^{-4} N south

5 a 1.50×10^{-3} N west **b** 3.00×10^{-3} N east

6 a 40.0 A into page **b** 20.0 A out of page

7 a 2.00×10^{-1} N m^{-1} west **b** 2.00×10^{-1} N m^{-1} east

8 a 2.0×10^{-3} N m^{-1} north-east **b** 1.0×10^{-3} N m^{-1} south-west

9 Magnetic flux due to wire N at point M is south.

10 a 2.00×10^{-9} N **b** 0 N **c** 2.50×10^{-4} N

3.3 Electric motors

1 1.00×10^{-1} N into page

2 1.00×10^{-1} N out of page

3 0 N as the field and current are parallel.

4 Anticlockwise

5 D

6 a Down **b** Up

7 Anticlockwise

8 a Down **b** Up **c** 0 N m

9 C

10 The commutator reverses the direction of the current through the coil of the motor at a particular point. This enables the resultant torque on the coil at that point to keep the motor rotating in a constant direction.

Chapter review

1 B

2 B

3 A

4 a R_4 **b** R_2 and R_3 **c** R_4 **d** E_2

5 a all four in series
b two in series that are connected to two in parallel
c all four in parallel
d one resistor that is connected to three in parallel

6 a 7.50×10^{-1} A, 1.20 A, 12.0 A, 2.25 A
b 7.50×10^{-1} A, 0.60 A, 3.00 A, 7.50×10^{-1} A

8 a 45.0 Ω **b** 1.80 W

9 20°C

10 a 75 V **b** 37.5 W

11 60, 60, 50, 0

12 a 0.23 A 5.7 V

13 a Y **b** 1.9×10^3 Ω **c** 40×10^{-3} A **d** 98 V **e** 3.9 W

14 a into the page **b** out of the page
c out of the page **d** out of the page

15 5.00×10^{-5} T south

16 into the page

17 1.00×10^{-4} T north

18 7.07×10^{-5} T

19 north-west
20 C
21 a 5.00×10^{-9} N into the page
b 2.00×10^{-3} N into the page
c 5.00×10^{-2} N into the page
22 a out of the page b into the page
23 0
24 a attraction b attraction c repulsion
25 0
26 2.00 N m^{-1}
27 0.10 N
28 anticlockwise
29 D

Chapter 4

4.1 Magnetic flux and induced currents

1 3.20×10^{-6} Wb, 2.26×10^{-6} Wb, 1.60×10^{-6} Wb, 0 Wb
2 -9.37×10^{-7} Wb, -1.60×10^{-6} Wb, -2.26×10^{-6} Wb, -3.20×10^{-6} Wb
3 a -3.20×10^{-6} Wb b -6.40×10^{-6} Wb
c 3.20×10^{-6} Wb d -1.60×10^{-6} Wb
4 a zero b negative c positive d negative
5 There must be a changing magnetic flux in the conductor that makes the coil, and the coil must be part of a complete circuit.
6 As S is closed a current in Y grows, which deflects the galvanometer needle to the right, and then drops to zero. While S is closed no current flows. As S is opened a larger current in Y grows, which deflects the galvanometer needle to the left, and then drops to zero.
7 As the current in X steadily decreases the current in Y is constant and deflects the galvanometer needle to the left. As the current in X steadily increases the current in Y is constant and deflects the galvanometer needle to the right.
8 a 1.01×10^{-5} Wb b zero
9 a -1.01×10^{-5} Wb b -1.01×10^{-2} Wb s^{-1} c 4.00 mA
10 a 8.00 mA from X to Y b 1.00 mA from X to Y
c 2.00 mA from X to Y

4.2 Induced EMF: Faraday's law

1 a 1.2×10^{-6} Wb b zero c 3.0×10^{-5} V d 2.0×10^{-5} A
2 a -8.00×10^{-5} Wb b 4.00×10^{-3} V c 2.00 V
3 a 1.00×10^{-4} Wb b -6.00×10^{-3} V
4 a 2.22×10^{-3} Wb b -1.78×10^{-2} V
5 D
6 5.00×10^{-2} T
7 C
8 a out of the page b into the page c out of the page
9 a positive b positive c negative

4.3 Electric power generation

1 a 1.00×10^{-5} Wb b 9.66×10^{-6} Wb c 8.66×10^{-6} Wb
d 7.07×10^{-6} Wb e 5.00×10^{-6} Wb f 2.590×10^{-6} Wb
g 0 Wb
2 -3.41×10^{-4} Wb s^{-1}, -9.99×10^{-4} Wb s^{-1}, -1.59×10^{-3} Wb s^{-1}, -2.07×10^{-3} Wb s^{-1}, -2.41×10^{-3} Wb s^{-1}, -2.59×10^{-3} Wb s^{-1}
3 It increases
4 3.41×10^{-2} V, 9.99×10^{-2} V, 1.59×10^{-1} V, 2.07×10^{-1} V, 2.41×10^{-1} V, 2.59×10^{-1} V
5 a 90° b 2.62 V
6 B
7 a C b D c C d B e D
8 0.900 T
9 a 339 V b 679 V c 3.39 A d 2.40 A
10 a 96.0 Ω b 339 V c 3.54 A

4.4 Transformers

1 a $V_p = -N_1 \frac{\Delta\Phi_B}{\Delta t}$ b $V_s = -N_2 \frac{\Delta\Phi_B}{\Delta t}$ c $\frac{V_p}{V_s} = \frac{N_1}{N_2}$
2 a A, B and D b A, B and D
3 a A b B and D
4 Power losses occur when electrical energy is converted into heat energy in the copper windings and in the iron core. Energy losses in the core are due to eddy currents.
5 a B b D c A
6 a 80.0 V b 0.200 A c 16.0 W
7 a 40.0 turns b 0.100 A c 24.0 W
8 No. In order for an EMF to be generated in the secondary coil a changing magnetic flux is required, a constant DC supply will create a constant field, therefore no EMF is induced in the secondary coil.
9 1200 J

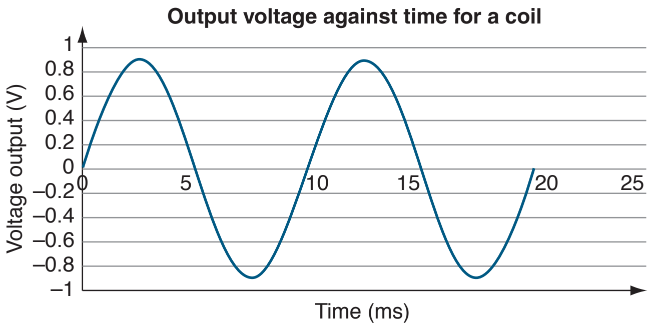
4.5 Distributing electricity

1 a By transforming to higher voltages the current decreases, this allows thinner cables to be used. Also the power lost by the cables is reduced significantly as $P_{loss} = I^2R$
b The 'corona effect' limits the voltage for power transmission. Differences in potential of 1000 V per cm will cause current to flow through air. At 500 kV, any ground source must be at least 500 cm away from the active wire.
2 a 2.0×10^3 A b 1.00×10^3 A
3 a 8.00% b 2.00%
4 1.00%
5 a 10.0 A b 4.0×10^2 W c 8.00% d 4.60×10^2 V
6 a 1.00 A b 0.0800% c 4.996×10^3 V
7 a \$0.28 b \$0.0056 c \$0.42 d \$0.14 e \$3.68
8 a 4.00×10^6 V, no b 1.10×10^5 V c 2.50×10^7 W
9 a 15.0 A b 9.97×10^3 V c 4.50×10^2 W, no
10 a 150.0 A b 7.00×10^2 V c 1.05×10^5 W d no

Chapter review

1 a 3.20×10^{-3} A clockwise b 3.20×10^{-3} A anticlockwise
c 1.60×10^{-3} A anticlockwise
2 a 2.01×10^{-3} V b 2.01×10^{-3} A
3 a A b C c B
4 a To the left, the soft iron core is induced to become a temporary magnet
b To the left, attraction c to the right, repulsion
5 2.01×10^{-2} A from Y to X
6 A and C
7 from X to Y
8 a 4.00×10^{-3} A from X to Y b 8.00×10^{-6} N to the left
9 No current as switch is open
10 5.00×10^{-2} N
11 to the right
12 1.00×10^{-2} N
13 to the left
14 a 1.60×10^{-3} V b 0
15 5.00×10^{-5} A
16 a 5.00×10^{-6} Wb b 0 Wb c 2.50×10^{-3} V
d 1.25×10^{-3} A e no
17 a 2.00×10^{-4} A b 2.12×10^{-1} T

18 a



b 0.636 V
c halved to 5 ms, V_{peak} doubles to 1.8 V, V_{RMS} becomes 1.3 V

19 a 1.00 A b 10 turns c 42.0 W

20 a C b A c B

21 C

22 C

23 a 5.0×10^2 Hz b 17.7 V c 50 V

24 a 188 W b 375 W c 1.67 Ω

25 C

26 C

27 a 8.91 V b 17.8 V c 89.9 Hz d 9.45 V

28 a 2.00 A, 236.0 V, yes
b Dropping the voltage increases current and results in power loss c 0.267 d 800.0 W

29 step-down transformer, 239 V, 239 V, 237 V

30 Yes

The sounds of music

E1 1.58×10^{-8} W m^{-2}

E2 2.27×10^{-3} W m^{-2}

E3 $\sqrt{3} \times f_1$

E4 50.0 Hz

E5 a 1.73 m b 0.867 m

E6 1.28×10^3 Hz

E7 6 m

E8 4.98 m

E9 constant temperature ensures a consistent sound

E10 loosen the string

E11 11.1%

E12 2.01%

E13 loosen or tighten the string until the point frequency of the tuning fork and violin string are the same

E14 B

E15 C

E16 3.16 A

Chapter 5

5.1 Wave properties

1 Both involve the transfer of energy from one place to another without the net movement of particles.

2 Mechanical waves involve the physical transfer of vibration from one particle to another. The bonds between solid particles are more rigid and closer together so the energy will be transferred more rapidly.

3 160 Hz, 6.25×10^{-3} s

4 1.8 m s^{-1}

5 Decrease the frequency, as this would allow more time between each wave, therefore increasing the length of each wave.

6 C

7 C and E

8 A

9 B

10 a Sound energy to kinetic energy of the microphone to electrical energy of the signal.
b Maximum pressure variation occurs at: 0.50 s, 1.50 s, 2.50 s, 3.50 s, 4.50 s and 5.50 s.

5.2 Wave behaviour

1 a 3.4×10^{-1} m b 3.4×10^{-1} m c 340 m s^{-1}

2 D

3 a 1.70×10^{-1} m b 1.70×10^{-1} m c 340 m s^{-1}

4 C

5 343 m s^{-1}, 3.90×10^{-1} m

6 349 m s^{-1}, 3.97×10^{-1} m

7 51.2°

8 79.4°

9 a 3.40×10^{-1} m
b Since the aperture is approximately the same as the wavelength appreciable diffraction will occur, resulting in significant sound energy arriving at positions P and R

10 a 8.50×10^{-2} m
b Since the aperture is greater than the wavelength less diffraction will occur, resulting in less sound energy arriving at positions P and R
c Since less diffraction occurs, more sound energy arrives at point Q

5.3 Wave interactions

1 a true b false c true d false

2 When the glass is exposed to sound of the same frequency as its natural frequency of vibration, resonance will occur. The amplitude of the vibrations then will increase if sufficient energy is supplied, and the glass will shatter.

3 The sound box of a guitar is tuned to resonate in the range of frequencies being produced by the guitar strings. Resonance within the sounding box amplifies the sound.

4 As a result of the superposition of two waves of equal amplitude and frequency travelling in opposite directions in the same medium.

5 a C b B c C

6 a at the centre
b $^1/_4$ of the way along the string c $^1/_6$ of the way along the string

7 a 3.00×10^2 Hz b 6.00×10^2 Hz c 9.00×10^2 Hz

9 a 9.00×10^{-1} m b 4.50×10^{-1} m c 1.10×10^3 Hz

10 a 1.10×10^2 Hz b 3.30×10^2 Hz c 7.70×10^2 Hz, 5.50×10^2 Hz

5.4 Electromagnetic radiation

1 a any three of: radio waves, infrared, visible light, ultraviolet, X-ray
b speed in a vacuum, electrical and magnetic field components

2 a 5.36×10^{-7} m b 3.71×10^{-19} J, 2.32 eV

3 A–E

4 a i 1.95 eV ii 3.12×10^{-19} J b 8.28×10^5 m s^{-1}

5 C

6 a false b true c false d true

7 E
8 B and E
9 **a** 4.60×10^{14} Hz **b** 1.90 eV **c** 2.00 eV
d 0.0851 eV **e** 1.58×10^{-25} kg m s^{-1}
10 **a** -8.51×10^{-2} V
b The photoelectrons do not have sufficient kinetic energy to reach the anode.
c −0.236 V

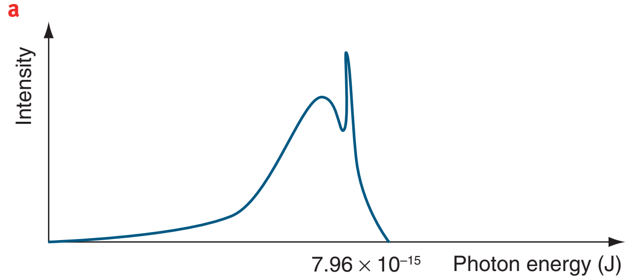
5.5 Electromagetic radiation and matter

1 **a** 1.50×10^{-6} m **b** D
2 **a** 2.46×10^{15} Hz **b** 1.03×10^{-7} m **c** 13.6 eV
3 **a** 12.1 eV **b** no
c It would eject the electron causing the atom to become ionised. The ejected electron would have (14.0 − 13.6) = 0.40 eV of kinetic energy.
4 **a** 12.7 eV
b 12.72 eV, 2.51 eV, 0.63 eV, 12.09 eV, 1.88 eV, 10.21 eV
5 Any excess energy above the ionisation energy is retained by the ejected electron in the form of kinetic energy, which may be any value, as kinetic energy is not quantised.
6 3.90×10^{-7} m
7 10
8 Emit more; emitting light will result from transitions from higher energy levels to any lower energy level including the ground state, however, absorption can only occur from the ground state to a higher level
9 no
10 De Broglie proposed a model of the atom in which the electrons were viewed as matter waves with resonant wavelengths that determined the circumference of the energy levels. This is similar to the resonant wavelengths that fit the length of a violin string.

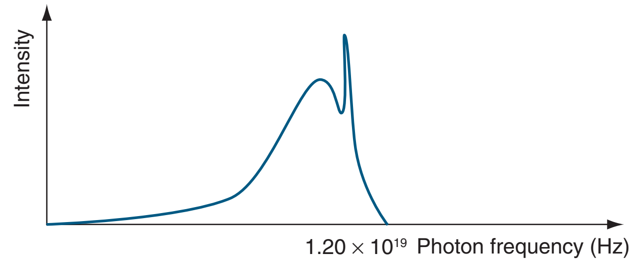
5.6 X-rays

1 **a** 1.12×10^{-14} J **b** 1.57×10^{8} m s^{-1}
c 70.0×10^{3} eV or 1.12×10^{-14} J
2 **a** 150×10^{3} eV or 2.40×10^{-14} J
b 3.62×10^{19} Hz **c** 8.29×10^{-12} m
3 **a** 3.55×10^{-11} m
b This is the wavelength of the most energetic photons that can be produced by the most energetic bombarding electrons. A smaller wavelength would have to come from a more energetic bombarding electron, which cannot be produced by this accelerating potential difference.
4 8.45×10^{18} Hz
5 The two peaks in the spectrum are the line spectrum caused by the bombarding electrons ejecting inner orbital electrons from the molybdenum atoms. When these inner electrons are ejected they are replaced by outer electrons in transitions that produce high energy X-ray photons.
6 **a** 4.01×10^{4} eV
b The peak is due to the bombarding electrons ejecting inner orbital electrons from the barium atoms. The inner electrons are then replaced by outer electrons in transitions that produce characteristic high energy X-ray photons of a particular wavelength due to the specific energy transition that occurs in barium.
c This is the wavelength of the most energetic photons that can be produced by the most energetic bombarding electrons. A smaller wavelength would have to come from a more energetic bombarding electron.
d 4.97×10^{4} eV

7 **a**



b

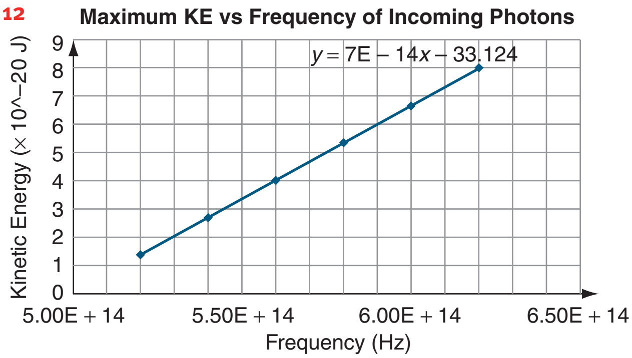


8 Tungsten is used because it has a very high melting point temperature. Copper is used because it is a very good conductor of electricity and has a very high conductivity of heat.
9 B
10 An X-ray tube produces X-rays in a single burst; a synchrotron can generate them continuously for hours. X-rays from an X-ray tube can pass through lighter atoms; those produced in a synchrotron are more likely to interact with these atoms. Synchrotron X-rays are 100 million times brighter than X-rays from traditional sources.

Chapter review

1 4.0 m
2 A
3 A
4 C
5 C and D
6 1.02×10^{2} Hz
7 3.05×10^{2} Hz
8 **a** **i** 3.43×10^{-19} J **ii** 2.14 eV
b 1.14×10^{-27} kg m s^{-1} **c** 1.46×10^{21} photons
9 **a** 4.60×10^{-19} J **b** 1.53×10^{-27} kg m s^{-1}
c 5.0×10^{20} photons **d** 2.30×10^{-3} W
10 A
11 **a** The wave model predicts that light of any frequency will emit photoelectrons from a metallic surface if given sufficient time.
b The wave model predicts that the energy delivered to the electrons by a light beam of constant intensity will be proportional to time. This suggests that a low intensity light beam should take longer to eject photoelectrons from a metallic surface.
c According to the wave model of light, a higher intensity beam will deliver more energy to the electrons and will therefore produce photoelectrons with higher kinetic energy than a lower intensity beam.

12



a 6.7×10^{-34} J s **b** Planck's constant **c** 5.0×10^{14} Hz **d** no

13 **a** 2.07 eV **b** 39.8×10^{-20} J **c** 2.69×10^{-25} kg m s^{-1}
d 2.49×10^{-1} V

14 C

15 $f_{3-1} = 1.62 \times 10^{15}$ Hz, $f_{3-2} = 4.34 \times 10^{14}$ Hz,
$f_{2-1} = 1.19 \times 10^{15}$ Hz

16 there is no shortest wavelength

17 When the excited electron returns to the ground state, it emits a photon of the same energy as it absorbed but the direction is random. Some photons travel in the same direction as the incoming photons but when resolved in the absorption spectrum are far less intense than the surrounding photons.

18 **a** 9.05×10^{20} photons **b** infrared

19 **a** 2.90×10^{21} photons **b** 1000 W

20 B

21 C

22 A

23 8.019×10^{-19} J or 5.01 eV

24 Only certain frequencies of light will cause photoelectrons to be emitted from a surface; there is no time delay between the absorption of photons of different intensities and the emission of a photoelectron; the maximum kinetic energy of the ejected photoelectrons is the same for different light intensities of the same frequency.

25 **a** 2.12×10^{15} Hz, 1.41×10^{-7} m
b 1.18×10^{15} Hz, 2.54×10^{-7} m
c 5.07×10^{14} Hz, 5.92×10^{-7} m

26 **a** 65.0 eV or 1.04×10^{-17} J
b 4.78×10^{6} m s^{-1}
c 1.52×10^{-10} m

27 **a** 1.41×10^{-18} J **b** 1.41×10^{-7} m **c** 10.4 eV or 1.66×10^{-18} J

28 **a** 1.12×10^{-18} J **b** $n = 3$
c 0.78×10^{-18} J or 4.90 eV, 1.07×10^{-18} J or 6.70 eV, 2.88×10^{-19} J or 1.80 eV
d 1.86×10^{-7} m

29 3.05×10^{-8} m

30 **a** no
b This is sufficient energy to ionise the mercury atoms. The liberated electrons are then free to conduct current along the tube.
c 2.54×10^{-7} m

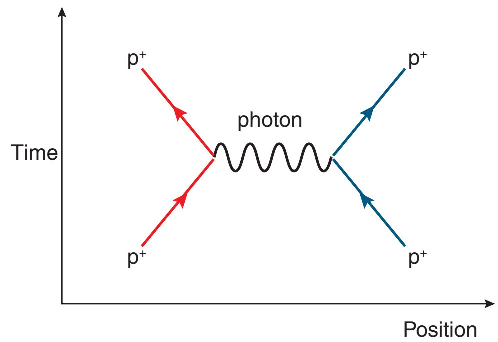
Chapter 6

6.1 Extending our model of matter

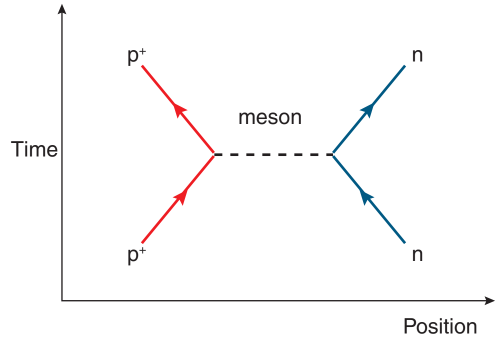
1 Strong nuclear 10^{38} : electromagnetic 10^{36} : weak nuclear 10^{32} : gravitational 1

2 **a** 5.00×10^{-7} m **b** 1.00×10^{-11} m **c** 6.265×10^{-15} m

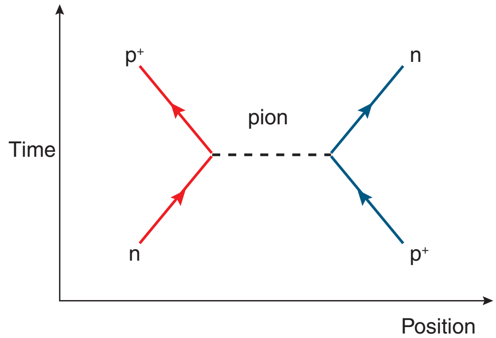
3 **a**



b



c



4 **a** 3.0107×10^{-10} J **b** 3.0148×10^{-10} J

5 **a** 1.40×10^{-14} J each particle **b** 1.50×10^{-16} J each particle
c 1.34×10^{-11} J each particle

6 **a** down, down, up **b** anti-up, anti-up, anti-down
c up down strange **d** charm, anti-down

7 **a** proton **b** rho-minus **c** kaon-minus **d** sigma-minus

6.2 Einstein's special theory of relativity

1 Galileo realised that a force changes motion, not causes it.

2 **a** 0 m s^{-1} **b** 2.99×10^{4} m s^{-1}
c The person at the equator has acceleration towards the centre of the Earth; both people also have an acceleration towards the Sun.

3 24.2% difference

4 The GPS picks up its signal from a satellite that is stationary in the Earth's frame of reference.

5 **a** 376 m s^{-1} **b** 306 m s^{-1} **c** 346 m s^{-1} **d** 376 m s^{-1}

6 A

7 **a** If there was an apparent force on you towards one of the walls of the room, or if a small pendulum maintained an angle to the vertical.
b The motion of the merry-go-round is not constant as it is changing direction constantly, it is accelerating towards the centre of the circular path.

8 **a** –15.0 m s^{-1} **b** 3.00 m **c** 0.200 s

9 **a** 50.0 m s^{-1} **b** 50.0 m s^{-1}

10 A

6.3 To the stars

1 3600 arcsec
2 1800 arcsec
3 1×10^3 m
4 Hipparchus called the brightest stars 'first magnitude' (+1) and the dimmest 'sixth magnitude' (+6), brighter stars discovered after Hipparchus had to then have negative values relative to Hipparchus's brightest.
5 The physics of heat transfer from hot bodies was well known, so accurate predictions could be made.
6 Nuclear forces are 10^4 times greater than electromagnetic forces, therefore there are 10^8 times more energy produced in nuclear reactions than chemical reactions.
7 Assumptions about the mass, temperature, heat content, ability to transfer heat, and the pressure gradient and more.
8 Fusion occurs in the inner $0.25R$ of the Sun, energy flows from this region by radiative diffusion and then convection. EM radiation 'bounces' from particle to particle, transferring heat energy in the process out to about $0.7R$. Then hot gases rise from this region to the upper regions by convection. The hot gases on the surface radiate their energy out into space by EM radiation, including in a direction towards the Earth.
9 C
10 A

6.4 Fundamentals of astronomy

1 C
2 In Brisbane the SCP would be 27° above the horizon in the south, while in Perth the SCP is 32° above the horizon in the south.
3 **a** C **b** D
4 **a** Pollux **b** Formalhaut
5 Stars appear move 1° further from their position at the same time in each subsequent night. Orion will be 7° further west of due north after one week.
6 Aquarius rose first, about 7 hours before Orion.
7 C
8 **a** It won't be able to be seen as it will be in line with the Sun and Earth.
b It is in a position known as 'opposition' only the superior planets can be seen in this position.
9 **a** 102.1%, 100.1%, 100.5%
b 2.1×10^{-3} m, 1×10^{-4} m, 5×10^{-4} m
10 **a** 2.83 Earth years **b** 4.00 AU

6.5 Hubble's universe

1 About 15 times the Milky Way galaxy away; it would have an angular diameter of about 5°; it is too far away so it is very faint.
2 The faint Cepheid is about twice the distance away than the brighter Cepheid.
3 **a** 1.96×10^{12} pc³ **b** 19.6 pc³ per star **c** 3.35 pc
d The nearest stars to the sun are about 1 pc away. The figure calculated above is an average distance as the inner stars are closer while the outer stars are sparser.
4 3.0×10^7 m s⁻¹
5 6.3×10^6 m s⁻¹; 2.1% of the speed of light
6 By looking at distant objects astronomers are looking far back in time to the early Universe. We see distant objects as they were earlier in their life-cycle.
7 Steady state theory: the Universe is infinite and eternal. Big Bang: the Universe is expanding. The steady state theory requires that matter is being continuously created; experiments on the large scales of distance and time that characterise the Universe are extremely difficult to create.
8 There is no way to tell.
9 The wavelength of the radiation has increased due to the expansion of space.
10 This value leads to an age of the Universe of about 2 billion years. This is less than the age of the Earth.

Chapter review

1 gluon, photon, W^+, W^- and Z^0, graviton
2 **a** 1.02×10^{-10} m **b** 1.53×10^6 m
3 1.84×10^{-13} J
4 Bosons – mediate the strong nuclear, electromagnetic and weak nuclear forces; leptons – interact via the weak nuclear force, and if charged will interact via photons, each lepton has a corresponding neutrino; mesons – have a baryon number of zero, are made up of a quark and an anti-quark pair; baryons – have a baryon number of one, or negative one, if it is the antiparticle. Is made up of three quarks.
5 proton – up, up, down; neutron – up, down, down; anti-proton – anti-up, anti-up, anti-down; anti-neutron – anti-up, anti-down, anti-down.
6 A and C
7 C
8 A, B and C
9 A and C
10 B
11 B
12 A
13 When stopped at the station (i) and when travelling at constant speed (iii) there will be no difference, when accelerating (ii) you would feel a force acting on you from the train pushing you in the direction that the train is accelerating.
14 58°; Albany it would be lower and in Kununurra higher
15 **a** Arcturus **b** Betelgeuse **c** Alpha Centauri **d** Aldebaran
16 **a** RA 6 h 43 min, dec. −17°
b RA 1 h 36 min, dec. −58°
c RA 18 h 35 min, dec. +39°
d RA 5 h 12 min, dec. −8°
17 30.3 km s⁻¹
18 1.30×10^6
19 4.8 thousand years
20 95 billion years
21 The spectra from these stars show lines similar to the spectra from our Sun and known elements.
22 The period of the Cepheid variables is related to luminosity, so distance can be found, they are also very bright.
23 **a** 2.4 pc, no
b The distance from our Sun to its nearest neighbour is about 1 pc, this is less than half of the average distance between stars in the Milky Way.
24 temperature, elements and their state, pressure, magnetic field
25 34.6 m s⁻¹ or 124 km h⁻¹ towards us
26 4.4 km s⁻¹. These are very small velocities on a galactic scale.
27 Blueshifts imply that the galaxies are moving towards us, they must be moving towards us at a rate that is greater than rate at which the Universe is expanding.
28 The age of the Universe can also be found using the position on the H-R diagram and nuclear decay processes. All of these methods estimate an age of about 14 billion years.

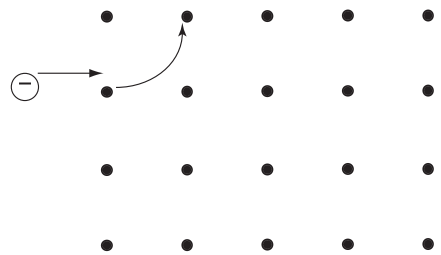
Chapter 7

7.1 Force on charges in magnetic fields

1 C
2 A
3 a south b C
4 A
5 a north b A
6 A particle with no charge, like a neutron, as it will not experience any force due to moving in the magnetic field, therefore its path will not change.
7 C and D
8 A and B as the particle must have the opposite charge, or have the field in the opposite direction, to experience a force in the opposite direction.
9 a south b i 2*F* ii 2*F* iii 4*F*
10 a 2*F* north b greater radius

7.2 Particle accelerators

1 B
2 a Electrons leave the hot cathode and accelerate towards a positively charged anode. They can be deflected as they pass through an electric field produced by oppositely charged parallel plates and a magnetic field generated by an electromagnet. They can be detected as they hit a fluorescent screen at the rear of the tube.
b The electrons are accelerated away from the negative cathode and towards a positively charged anode.
4 A circular accelerator, such as a cyclotron, can be used to accelerate particles within a more compact space than the equivalent operations of a very long linear accelerator.
5 a 5.93×10^7 m s^{-1} b 2.25×10^{-4} m
6 a



b The radius of the electron's path is dependent on its velocity and the magnitude of the magnetic field that is acting.
7 a 1.43×10^4 V m^{-1} b 9.52×10^6 m s^{-1}
8 2.96×10^7 m s^{-1}
9 9.41×10^{-4} T
10 a 9.63×10^{-15} N b 4.63×10^{-3} m

7.3 Synchrotrons

1 a i The linac consists of an electron gun, a vacuum system, focusing element and RF cavities.
ii Each time the charged particles travel around the circular booster ring they receive an additional energy burst from a radio-frequency (RF) chamber.
iii In the storage ring of the synchrotron, electrons revolve around for hours at a time at speeds close to the speed of light. A series of magnets make them bend in an arc as they travel through the ring. It is when the electrons change direction that they emit synchrotron radiation.
iv The beamline is the path taken by synchrotron light as it exits the storage ring towards an experimental station (or endstation).
b i The strength of the magnetic field is increased as the velocity of the electrons increases to account for energy losses due to the increasing effects of relativity which result from the increased velocity and relativistic mass by this stage.
ii Electrons move from a heated filament inside the electron gun within the linac.
iii RF cavities accelerate electrons within the linac.
iv Insertion devices are located in the straight sections of the storage ring.
2 a The precise configuration of the bending, focusing and steering magnets found in the storage ring.
b The specification of the lattice sets the parameters for the synchrotron light produced.
3 a Due to the presence of an oscillating electromagnetic field produced by the transformers extra energy is given to the electrons.
b The particles would gradually lose their energy through collisions with other atoms and the production of synchrotron light. Their orbital speed would be slowed and they would cease to produce synchrotron light.
c By replenishing the energy lost via a burst of energy from the RF cavity, the charged particles continue to move at the same speed in a path of constant radius in the storage ring. The particles of a cyclotron increase their energy with each revolution and so their radius of orbit increases each revolution.
4 To minimise energy losses through collisions between electrons and gas molecules.
5 B
6 An undulator consists of some hundred low-power magnetic poles aligned closely together in rows. The effect of the undulator is to deflect the electrons more gently, thus producing a much narrower beam of brighter synchrotron radiation that is enhanced at specific wavelengths. This is in contrast to a bending magnet, which produces a continuous broad cone of much less bright synchrotron light. The output from a wiggler is radiation that is not as bright as from the undulator, but brighter than that from the bending magnet. The wiggler forms a broad band of incoherent synchrotron light.
7 a 3 GeV
b Pharmaceutical development, mining and mineral exploration, manufacturing microstructures etc.
c Answers will vary.
8 a 2.05×10^8 m s^{-1} b no
9 a 1.14×10^{-3} m b no
10 Similarities: same basic design, a linac, booster ring, storage ring, number of beamlines and a third generation source. Differences: power output, Delta has fewer beamlines, and is half the size of the Australian Synchrotron. Delta is oval shaped while the Australian Synchrotron is symmetrical through all axes.

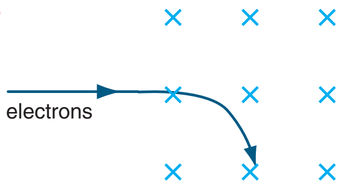
7.4 Mass spectrometry

1 2.61×10^{-5} m
2 1.22×10^{-2} m
3 2.72×10^{-26} kg

4 positive
5 The direction of the magnetic field would need to be reversed and the voltage used to accelerate the ions would need to be reversed too.
6 2.09×10^4 V
7 3.76×10^{-2} T
8 no

Chapter review

1 B
2 The radius of curvature of both electron beams is the same therefore the velocity of both beams must be equivalent Using the right-hand rule $\boldsymbol{B}_x$ must be out of the page for the electron beam to bend up the page, and $\boldsymbol{B}_y$ must be into the page for the electron beam to bend down the page.
3 C
4 A
5 2.00×10^{-14} N
6 B
7 9.92×10^7 m s^{-1}
8 1.40×10^5 V m^{-1}
9 D
10 1.99×10^{-5} m
11 2.93×10^{-15} J, 1.83×10^4 eV
12 A
13 2.02×10^{-16} J
14



15 5.81×10^{-2} m
16 1.58×10^6 m s^{-1}
17 5.17×10^{-1} m
18 6.62×10^{-27} kg
19 2.72×10^{-26} kg

Index

Page numbers in **bold** refer to key terms in **bold** type in the text
Entries marked 'CD' refer to the Student CD

A

B

C

D

E

F

G

N

S

T

U

V

W

X

Z