



heinemann

physics 12

3rd edition

Rob Chapman
Keith Burrows
Carmel Fry
Doug Bail
Alex Mazzolini
Jacinta Devlin
Henry Gersh

PEARSON
Heinemann

VCE Units 3 and 4

Pearson Heinemann

An imprint of Pearson Education Australia
A division of Pearson Australia Group Pty Ltd
20 Thackray Road, Port Melbourne, Victoria 3207
PO Box 460, Port Melbourne, Victoria 3207
www.pearsoned.com.au/schools

Other offices in Sydney, Brisbane, Perth and Adelaide and associated companies throughout the world.

Copyright © Doug Bail, Keith Burrows, Robert Chapman, Carmel Fry, Alex Mazzolini, Geoff Millar 2009

First published 2009 by Pearson Education Australia
(a division of Pearson Australia Group Pty Ltd)

2012 2011 2010 2009

10 9 8 7 6 5 4 3 2 1

Reproduction and Communication for educational purposes

The Australian Copyright Act 1968 (the Act) allows a maximum of one chapter or 10% of the pages of this work, whichever is the greater, to be reproduced and/or communicated by any educational institution for its educational purposes provided that the educational institution (or the body that administers it) has given remuneration notice(s) to Copyright Agency Limited (CAL) under the Act. For details of the CAL licence for educational institutions contact Copyright Agency Limited (www.copyright.com.au).

Reproduction and Communication for other purposes

Except as permitted under the Act (for example a fair dealing for the purposes of study, research, criticism or review) no part of this book may be reproduced, stored in a retrieval system, communicated or transmitted in any form or by any means without prior written permission. All inquiries should be made to the publisher at the address above.

This book is not to be treated as a blackline master; that is any photocopying beyond fair dealing requires prior written permission.

Publisher: Malcolm Parsons
Editor: Catherine Greenwood
Text Designer: Rebecca Harrison
Cover Designer: Rebecca Harrison
Copyright and Pictures Editor: Katherine Wynne, Jacqui Liggett
Project Editor: Jane Sunderland
Production Controller: Jem Wolfenden
Illustrators: Guy Holt, Margaret Hastie
Typeset in 9.5pt Palatino by Palmer Higgs
Printed in China

National Library of Australia Cataloguing-in-Publication entry

Heinemann physics 12 / Rob Chapman ... [et al.].

3rd ed.

9781740819381 (pbk.)

Includes index.

For secondary school age.

Physics--Textbooks.

Other Authors/Contributors: Chapman, Rob.

Dewey Number: 530

Pearson Australia Group Pty Ltd ABN 40 004 245 943

Disclaimer/s

The selection of Internet addresses (URLs) provided for this book were valid at the time of publication and chosen as being appropriate for use as a secondary education research tool. However, due to the dynamic nature of the Internet, some addresses may have changed, may have ceased to exist since publication, or may inadvertently link to sites with content that could be considered offensive or inappropriate. While the authors and publisher regret any inconvenience this may cause readers, no responsibility for any such changes or unforeseeable errors can be accepted by either the authors or the publisher.

Acknowledgements

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

© Fundamental Photographs / Ken Kay: pp. 415l, 416.

© Mark Horsburgh/Reuters/Picture Media: p. 36.

© STFC: pp. 463, 464.

Aaron Gold / About.com:Cars: p. 266.

Alamy Limited: pp. 162, 234t, 275r, 276tr.

Alex Mazzolini: pp. 526, 536, 538.

All Australian Nature & General PhotoLibrary: p. 349l.

Australian Associated Press Pty Ltd: pp. 6b, 32, 37r, 40 (boxing), 66, 70.

Australian Synchrotron: pp. 476, 482, 495.

CERN: p. 471.

Connex, A Veolia Transport Company: p. 353.

Corbis Australia Pty Ltd: pp. 35l, 48bl, 48c, 57l, 65, 187, 197b, 406b, 491, 568, 569.

Courtesy of Brookhaven National Laboratory: p. 465t.

Dale Mann: p. 551.

Dorling Kindersley: pp. 264r, 359.

Doug Bail: pp. 156t, 541.

Dr Richard Porcas / Max Plank Institute for Radiostromie, Bonn: Germany: p. 258.

Dr Dave Irvine-Halliday: p. 166.

Flickr / Abraham Orozco: pp. 60; Jeff Medaugh: p. 28.

Getty Images Australia Pty Ltd: pp. 1, 10, 16, 33l, 35r, 40 (sumo), 186, 200, 226t, 236, 241, 243r, 250l, 549,

Institute Fumikrotechnik Maine Germany: p. 490.

iStockphoto: pp. 13b, 25, 35c, 37l, 40 (tennis), 48br, 48t, 57r, 64, 72, 125, 134, 137r, 147, 159, 160t, 164, 165, 226b, 234bl, 235, 240, 250r, 257, 264l, 269, 272, 274bl, 274r, 282 (computer) (stereo) (television) (laptop), 283, 285 (drill) (light) (torch) (keyboard), 289b, 389t, 403b, 457t, 508, 521t, 550, 557, 575, 577, 590.

Karl Ludwig, Yiyi Wang and Ahmet Ozcan: p. 502.

Keith Burrows: pp. 286, 288, 289t, 289c, 290, 293, 295, 307, 309, 312bl, 312tl, 312r, 315, 316, 317, 320, 323, 333t, 349r.

Malcolm Cross: pp. 150, 156b, 167, 171, 173, 174, 175, 176, 329t, 375, 382 410bl, 410br.

Mark Fergus: p. 53.

NASA: pp. 14b, 41, 77, 91, 97, 102, 108, 189, 347, 401, 511r, 512.

Northside Productions: p. 299.

PEA / Alice McBroom: pp. 161, 448t; Jacqui Liggett: p. 264c; Sarah James: p. 403t.

Peter Colman: p. 496.

Peter O'Donoghue: p. 312br.

Photodisc: pp. 224, 343, 376b, 465b, 466, 489, 585.

Photolibrary Pty Ltd: pp. 2, 13t, 90, 94, 151, 153, 197t, 204, 233, 251, 267, 276bl, 327, 346, 360, 369, 370, 376t, 379, 384, 390, 391, 402, 411, 415r, 436, 437, 442, 443, 447, 448b, 452, 469, 488, 493, 494, 507, 511l, 513, 523r, 544, 545, 567.

Roaring 40s: p. 325.

Shutterstock: pp. 14t, 21, 33r, 40 (bowling), 115, 116, 137l, 160b, 185, 237, 243l, 274tl, 275bl, 275tl, 276br, 276tc, 276tl, 281, 282 (circuit board), 285 (guitar), 326, 329b, 358, 457b, 521b, 523l, 573, 578, 582, 592, 593.

Siemens Press Picture: p. 389b.

The Picture Source / Terry Oakley: pp. 6t, 23, 333b, 406tl, 444.

Wikipedia / John Lanoue: pp. 82; Stevage: p. 234br; Andreas Frank: p. 516;

Harout S Hedeshian: p. 528.

Every effort has been made to trace and acknowledge copyright. However: should any infringement have occurred: the publishers tender their apologies and invite copyright owners to contact them.

The
publisher's
policy is to use
paper manufactured
from sustainable forests

CONTENTS

Introduction	vi
About the authors	vii

Unit 3

Area of study 1 Motion in one and two dimensions

Chapter 1 Motion	2
1.1 Mechanics review	3
1.2 Newton's laws of motion	13
1.3 The normal force and inclined planes	18
1.4 Projectile motion	23
Chapter review	30

Chapter 2 Collisions and circular motion	32
---	-----------

2.1 Momentum and impulse	33
2.2 Conservation of momentum	39
2.3 Work, energy and power	44
2.4 Hooke's law and elastic potential energy	51
2.5 Circular motion	57
2.6 Aspects of horizontal circular motion	62
2.7 Circular motion in a vertical plane	68
Chapter review	75

Chapter 3 Gravity and satellites	77
---	-----------

3.1 Newton's law of universal gravitation	78
3.2 Gravitational fields	84
3.3 Satellites in orbit	90
3.4 Energy changes in gravitational fields	99
3.5 Apparent weight and weightlessness	105
Chapter review	110
Exam-style questions—Motion in one and two dimensions	112

Area of study 2 Electronics and photonics

Chapter 4 Electronics	116
4.1 Analysing electronic circuits	117
4.2 Diodes	129
4.3 Amplification	137
Chapter review	144

Chapter 5 Introducing photonics	147
--	------------

5.1 Photonics in telecommunications	148
5.2 Optical transducers	155
5.3 Audio transmission via a light beam	171
Chapter review	178
Exam-style questions—Electronics and photonics	179

Detailed studies

Chapter 6 Einstein's special relativity	186
--	------------

6.1 Two principles Einstein did not want to give up	187
6.2 Einstein's crazy idea	197
6.3 Time is not what it seems	207
6.4 Time and space	215
6.5 Momentum, energy and $E = mc^2$	221
Chapter review	230

Chapter 7 Materials and their use in structures 233

7.1	External forces acting on materials	234
7.2	Stress and strength	239
7.3	Strain	245
7.4	Young's modulus	248
7.5	Strain energy and toughness	254
7.6	Forces in balance: translational equilibrium	260
7.7	Torque	264
7.8	Structures in translational and rotational equilibrium	269
	Chapter review	279

Chapter 8 Further electronics 281

8.1	Principles and practicalities of electronic design	282
8.2	Capacitors and time-varying circuits	293
8.3	Rectification and power supplies	299
8.4	Constructing and testing a working power supply	312
	Chapter review	321

Unit 4 Area of study 1 Electric power

Chapter 9 Magnets and electricity 326

9.1	Fundamentals of magnetism	327
9.2	The foundations of electromagnetism	333
9.3	Currents, forces and fields	337
9.4	Magnetic fields around currents, magnets and atoms	341
9.5	Forces on moving charges	346
9.6	Electric motors	351
	Chapter review	356

Chapter 10 Electromagnetic induction 358

10.1	Magnetic flux and induced currents	359
10.2	Induced EMF: Faraday's law	363
10.3	Direction of EMF: Lenz's law	368
10.4	Electric power generation	373
10.5	Alternating voltage and current	378
10.6	Transformers	482
10.7	Using electrical energy	386
	Chapter review	394
	Exam-style questions—Electric power	397

Area of study 2 Interactions of light and matter

Chapter 11 The nature of light 402

11.1	Review of light and waves	403
11.2	The wave model established	410
11.3	Photoelectric effect: Counterevidence for wave model	421
11.4	The dual nature of light	428
	Chapter review	434

Chapter 12 The nature of matter 436

12.1	Matter waves	437
12.2	Photons shed light on atom structure	446
12.3	Bohr, de Broglie and standing waves	454
	Chapter review	459
	Exam-style questions—Interactions of light and matter	460

Area of study 3 Detailed studies

Chapter 13 Synchrotron and applications 464

13.1	Particle accelerators	465
13.2	Synchrotrons	473
13.3	Synchrotron radiation	486
13.4	Scattering and beyond	498
	Chapter review	505

Chapter 14 Photonics 507

14.1	Incoherent light sources	508
14.2	Coherent light sources: lasers	518
14.3	Optical fibres	523
14.4	Applications of optical fibres	538
	Chapter review	547

Chapter 15 Sound 549

15.1	The nature of sound	550
15.2	The wave equation	557
15.3	Diffraction of sound	561
15.4	Amplitude, intensity and the decibel scale	565
15.5	Frequency, perceived loudness and the phon	571
15.6	Making sound: strings and air columns	575
15.7	Recording and reproducing sound: the first and last links	587
	Chapter review	595

Solutions	597
Glossary	622
Index	628

heinemann physics 12

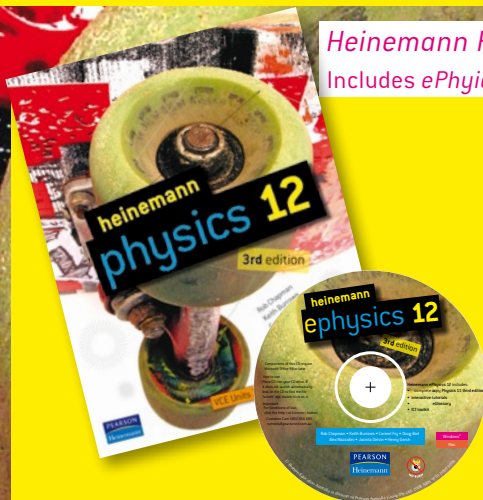
3rd edition

The complete package for Units 3 and 4 VCE Physics

Heinemann Physics 12 3rd edition is the most up to date and complete package for VCE Physics. The third edition has been fully revised and upgraded to match the content and focus of the new 2009 VCE Physics Study Design. Successful features of the second edition have been retained while significant improvements and innovations will make the books even easier and more and stimulating to use. *Heinemann Physics 11 3rd edition* covers Units 1 and 2 and *Heinemann Physics 12 3rd edition* covers Units 3 and 4.

Heinemann Physics 12 3e textbook

Includes *ePhysics* student CD



Key features:

- New full colour design
- All *detailed studies* in the textbook
- Extension and enrichment material clearly indicated
- Lesson-sized, self-contained sections
- Huge range of well-graded end-of-section questions and chapter reviews
- Exam-style questions that are exam style!
- Extensive glossary
- *ePhysics* interactive CD included with the text.

Each textbook includes *ePhysics*, an interactive student CD containing:

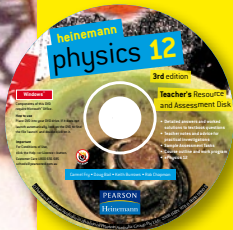
- Electronic textbook
- Interactive tutorials
- ICT Toolkit.

Heinemann Physics 12 3e Teacher's Resource and Assessment Disk

The Teacher's Resource and Assessment Disk contains a wealth of support material and makes effective implementation of the study design easy.

It includes:

- Detailed answers and worked solutions to all questions in the textbook.
- Extensive range of short and long practical activities all with teacher notes and suggested outcomes and answers
- Sample assessment tasks with marking guidelines
- Complete electronic copy of the textbook and *ePhysics*

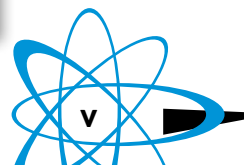


Heinemann Physics 12 3rd edition Companion Website
www.pearsoned.com.au/physics

The Companion Website includes further support for teachers including weblinks



www.pearsoned.com.au/physics



INTRODUCTION

The Heinemann Physics series is now in its third edition. The first edition was published in 1996 and since then the original author team have remained together and have continually striven to build on and improve the series. Over that time, not only has each remained highly involved in the teaching of physics, but they have also contributed to physics and physics education as members of professional organisations, supported curriculum development and regularly presented professional development to their colleagues.

The third editions of *Heinemann Physics 11* (Units 1 & 2) and *Heinemann Physics 12* (Units 3 & 4) represent the authors' ongoing commitment to physics teachers and students. The series has been fully revised and upgraded to match the content and focus of the new 2009 VCE Physics Study Design.

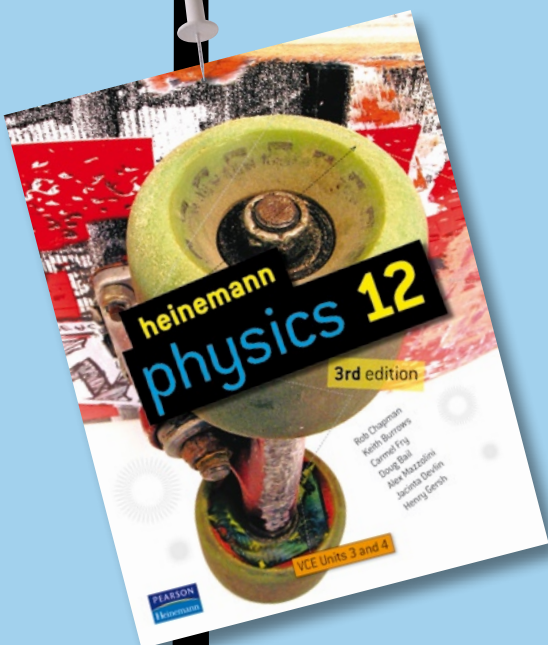
Successful features of the second edition have been retained while significant improvements and innovations have been added. These include:

- new full colour design
- all Detailed studies in the textbook
- exact match to structure and sequence of the study design
- chapters divided into student-friendly sections
- clear explanations and development of concepts consistent with the intent and scope of the study design
- extension and enrichment material clearly designated
- numerous well-graded end-of-section questions and chapter reviews
- exam-style questions that are exam-style!
- extensive glossary
- *ePhysics* interactive CD with each text.

The text 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 illustrated 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. At the end of each Area of study is a set of exam-style questions. These can be used for revision. The large number of questions is designed to access students' understanding of basic concepts as well as giving them practice at problem solving. Answers are supplied at the end of the text and extended answers and fully worked solutions are available on the Teacher's Resource and Assessment Disk.

Within each section, the concept development and worked examples occupy the main two-thirds column. The remaining on-third column has been set aside for some of the numerous 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.

The authors have written the text to follow the sequence, structure and scope of the Study Design. Material outside the scope of the Study Design is clearly marked. This includes entire sections and sub-sections.

This material has been included for a number of reasons, including as important background to core concepts, as important physics in its own right and as extension material for more able students. Teachers should consider whether they wish to incorporate this material into their work program.

The third edition includes all Detailed studies in the textbook. Three Detailed studies are available for each Unit. Students undertake one of these in each Unit. The Detailed studies for Unit 3 are Chapters 6–8 and the Detailed studies for Unit 4 are Chapters 13–15.

The textbook includes an interactive CD, *ePhysics 12*, which will enhance and extend the content of the texts. Included are:

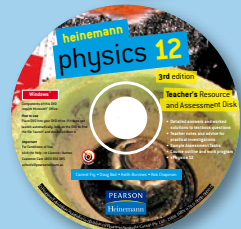
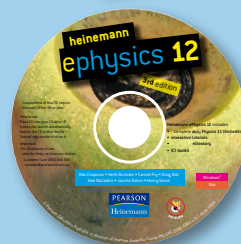
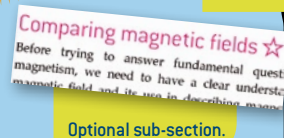
- fully interactive tutorials that allow students to explore important concepts which may be too difficult, dangerous or expensive to do first-hand in the classroom
- a complete electronic copy of the textbook
- ICT Toolkit with tutorials on spreadsheets, databases, web use and more.

The *Heinemann Physics 12 Teacher's Resource and Assessment Disk* supports the text and *ePhysics 12* and helps teachers implement, program and assess the course of study. Included are:

- detailed answers and worked solutions to all questions in the textbook
- extensive range of short and long practical activities, all with teacher notes and suggested outcomes and answers
- sample assessment tasks with marking guidelines
- link to the Companion Website pearsoned.com.au/physics
- complete electronic copy of the textbook and *ePhysics*.



The *Companion Website* (pearsoned.com.au/physics) includes further support for teachers, including weblinks.



Doug Bail

Is an experienced physics educator and writer with a particular interest in the development and integration of new technologies into science teaching. He has previously been a head of science and senior physics teacher and maintains a passion for making physics relevant, stimulating and accessible to all students. Doug now runs his own company developing and distributing products for physics education.

He led the development of the new practical activities which form part of the teacher support material. These activities were extensively trialled throughout Australia and include a range of activities from teacher demonstration to discovery-based investigation suiting a range of learning styles and needs. This includes many short activities for when time is limited!

Keith Burrows

Has been teaching senior physics in Victorian schools for many years. He is a member of the Australian Institute of Physics Victorian Education Committee and was actively involved with the VCAA in the design of the new course. Keith was a VCAA representative involved in introduction of the new VCE course to physics teachers in Victoria and running the workshop sessions for teachers. He is particularly keen to portray 'The Big Picture' of physics to students.

Rob Chapman

Has taught physics for many years from HSC onwards. Rob has been enthusiastic in exploring the possibilities presented by changing technologies over the years. He has been Science Coordinator at St Columba's College in Essendon, where he was instrumental in introducing the use of datalogging technology to junior science and senior physics classes. Rob is currently teaching senior physics at PEGS (Penleigh and Essendon Grammar School). He has written a wide variety of curriculum support material, including physics units for the CSFII. Rob has also produced physics trial examination papers and is the author of the acclaimed *Physics 12: A student guide*.

Carmel Fry

Has 19 years' involvement in development of text, CD and online curriculum materials for VCE physics and science. She is Head of Science at Ivanhoe Girls' Grammar School

where she continues her interest in providing high-quality curriculum resources and learning experiences for students. Carmel is the author of numerous texts, multimedia resources and teacher-resource materials developed for senior physics. These materials are currently in use in many parts of Australia and overseas. She has led the development of the Interactive Tutorials. Carmel is particularly passionate about providing physics curriculum materials that involve a variety of approaches to learning and that support independent learning through stimulating and appealing contexts and activities. Carmel would like to acknowledge the ongoing support of her husband and children over her many years of publishing.

Jacinta Devlin

Is an experienced teacher of science, maths and VCE physics. She has co-authored the Heinemann Science Links series for junior students and contributed to the development of support material including the VELs Teacher's Chronicle. Jacinta acknowledges and thanks Stefanie Pearce and Dr Mark Boland of the Australian Synchrotron Project for their assistance in researching and developing this unit. Jacinta's skills and experience have allowed her to make this cutting-edge topic not only exciting and relevant but also accessible to students of all backgrounds.

Review panel

The publisher and authors would like to acknowledge and thank the following people for their contribution to the text: the expert review panel consisting of experienced VCE teachers and educators—Luke Bohni, Mike Davies, Barry Homewood, Chris Hourigan, John Joosten, Terry Trevena, Steve Treadwell, Lyndon Webb and Chris Ward—and Dr Mark Boland, Accelerator Physicist, Australian Synchrotron, for his expert and generous input into the Synchrotron Detailed Study.

Acknowledgments

The publisher would like to acknowledge and thank the author team for their ongoing commitment and passion for this project. It is a huge and complex task and the demands, including short timelines, are great. Carmel, Keith, Rob, Doug and Jacinta, it has been a pleasure and privilege to work with you.

Unit 1

area of study 1

Motion in one and two dimensions

outcome

On completion of this area of study, you should be able to use the Newtonian model in one and two dimensions to describe and explain transport motion and related aspects of safety, and motion in space.

CHAPTER 1

Motion



Isaac Newton was born in England in 1642. He is considered to be one of the greatest scientific minds of all time, most famous for his laws of motion, but also for contributions in optics, astronomy and the nature of light and colour. He was driven by a desire to better understand the Universe, and he developed theories to explain its behaviour.

Newton studied at Cambridge, but returned to his birthplace of Woolsthorpe in 1665 when the plague swept through the cities of Europe. It was here that he realised that the force that keeps the Moon in its orbit around the Earth was the same force that caused objects to fall to the ground. His law of universal gravitation will be studied in Chapter 3.

Newton was a reclusive man and never married or had children. He is credited with some of the greatest scientific discoveries of all time. He developed his laws of motion, building upon work previously done by Galileo. These laws describe very precisely the behaviour of objects when forces act upon them, and they can be used with great confidence in everyday situations. However, for objects travelling at extreme speeds or in very strong gravitational fields, these laws are inaccurate.

Newton's use of mathematics and the scientific method of investigation revolutionised the study and development of scientific ideas. His most famous publication, the *Principia Mathematica*, was written in Latin and published in 1687. In this book, he demonstrated for the first time that complex motions as experienced by falling objects, projectiles and satellites can be explained and predicted by using a few simple rules. This is what you will be studying in the following chapters.

by the end of this chapter

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

- motion in one and two dimensions
- Newton's laws of motion
- inclined planes
- projectile motion.

1.1 Mechanics review

Before building on the mechanics that you learned in Year 11, it will be useful to revise the concepts and ideas that have been covered in the Year 11 course. This section is designed to provide a brief tour of the ideas and concepts that are used to describe and understand the motion of an object.

Vectors and scalars

Physical quantities can be divided into two distinct groups: *vectors* and *scalars*. These should always be written with an appropriate unit, for example 25 mm or 60 s.

- A *vector* quantity requires both a magnitude, or size, and a direction for it to be fully described. In mechanics, displacement, velocity, acceleration, force and momentum are common vector quantities.
- A *scalar* quantity requires only a magnitude for it to be completely described. Common scalar quantities include distance travelled, speed, time, mass and energy.

Vector techniques

Multiplying a vector by a scalar

When a vector is multiplied by a number (i.e. a scalar), the magnitude of the vector is changed by the appropriate factor but the direction remains the same. If the scalar is negative, this is interpreted as reversing the direction of the vector. For example, if $\mathbf{x} = 5$ m east, then $2\mathbf{x} = 10$ m east and $-\mathbf{x} = 5$ m west.

Vector addition and subtraction

Adding scalars is straightforward. The amounts are just added as one would add numbers. If a person walked 20 metres and then a further 30 metres, the total distance travelled is 50 metres, regardless of direction.

When adding *vectors*, seemingly unusual results are sometimes obtained; for example, $3 + 4 = 5$! This occurs because vectors have direction and are not always combined in one dimension. In adding two vectors, the tail of the second vector is placed at the head of the first vector. The sum of the vectors is the *resultant* vector. This is drawn as a directed line segment starting at the tail of the first vector and finishing at the head of the second vector. To describe the resultant, both the *magnitude* (given by the length of the vector) and its *direction* are required. General vectors are used in Figure 1.1 to illustrate this technique.

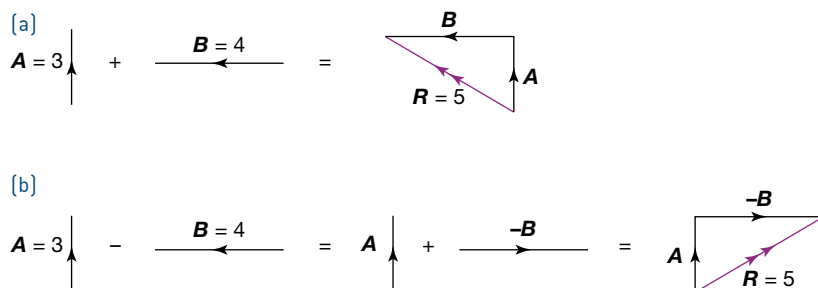


Figure 1.1 (a) The vector addition $\mathbf{A} + \mathbf{B}$ results in vector \mathbf{R} , drawn from the start of the first vector to the end of the second vector. (b) The vector subtraction $\mathbf{A} - \mathbf{B}$ results in vector \mathbf{R} , drawn from the start of the first vector to the end of the second vector. Pythagoras's theorem is used to find the size of the resultant vector.

Vector subtraction is required from time to time. This technique should be handled as a vector addition, but with the direction of the subtracted vector being reversed. For example, a change in velocity is given by:

$$\Delta \mathbf{v} = \mathbf{v} - \mathbf{u} = \mathbf{v} + (-\mathbf{u})$$

Describing motion

When a body moves from one place to another, its *displacement*, \mathbf{x} , is the vector given by the directed line segment from the start of the motion to the finish of the motion. Displacement is a vector, and indicates how far and in what direction the body has moved from the starting point. The displacement of a body may be different from the distance it has travelled. The displacement of a Year 12 student over a 24 h period might be zero, but he or she could have travelled many kilometres during the day. *Distance*, d , is a scalar quantity, and does not require a direction. The odometer in a car measures the distance travelled by the car.

Speed and *velocity* both indicate how fast a body is travelling, but there is an important difference between these quantities. The average speed, v , of a body is given by its rate of change of *distance*, whereas the average velocity, \mathbf{v} , of a body is its rate of change of *displacement*. The SI unit for both quantities is metres per second, m s^{-1} , but speed is a scalar quantity and velocity is a vector quantity.

$$\text{Average speed } v = \frac{\text{distance travelled}}{\text{time taken}} = \frac{d}{\Delta t}$$

$$\text{Average velocity } \mathbf{v} = \frac{\text{displacement}}{\text{time taken}} = \frac{\mathbf{x}}{\Delta t}$$

An *instantaneous velocity* or *speed* is the velocity or speed of a body at a particular instant in time. The speedometer in a car gives the instantaneous speed of the car.

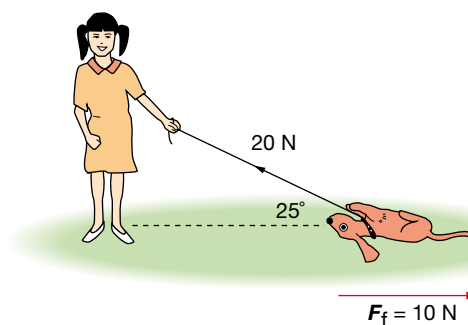
The average *acceleration* of a body is given by the rate of change of velocity over a given time interval. For example, if a car speeds up from 20 km h^{-1} to 60 km h^{-1} in 4.0 s , its average acceleration is 10 km h^{-1} per second. Acceleration is a vector quantity, its SI unit being metres per second squared (m s^{-2}). The symbols \mathbf{u} and \mathbf{v} are used to denote the initial velocity and the final velocity of a body over a given time interval.

$$\text{Average acceleration is given by } \mathbf{a} = \frac{\Delta \mathbf{v}}{\Delta t} = \frac{\mathbf{v} - \mathbf{u}}{\Delta t}$$

where $\Delta \mathbf{v}$ is the change in velocity over the time interval Δt .

Worked example 1.1A

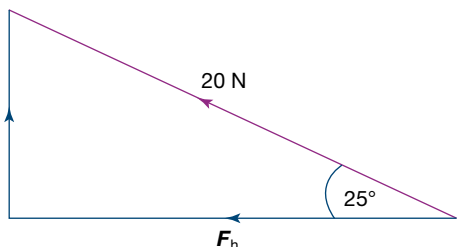
Rebecca is pulling her toy dog along the lawn with a force of 20 N applied at an angle of 25° to the horizontal. There is a frictional force of 10 N acting on the toy dog. Determine the net horizontal force acting on the dog.



Solution

The horizontal component of the pulling force is: $F_h = 20\cos 25^\circ = 18.1$ N. The net horizontal force is:

$$18.1 - 10 = 8.1 \text{ N}$$



Worked example 1.1B

A white cue ball travels over a pool table at 3.0 m s^{-1} east. It collides with the cushion and rebounds at 2.0 m s^{-1} west. If the collision time with the cushion is 25 ms, calculate:

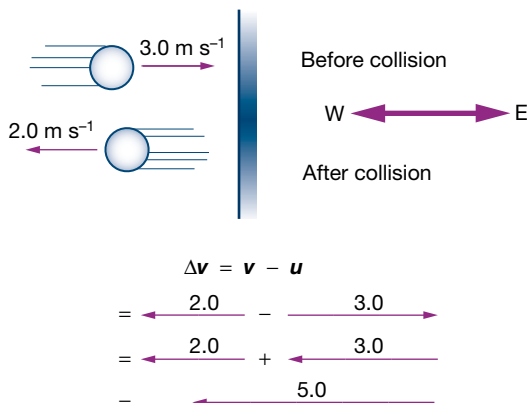
- the change in speed of the ball
- the change in velocity of the ball
- the average acceleration of the ball during its collision with the cushion.

Solution

- a** The white ball has slowed from 3.0 m s^{-1} to 2.0 m s^{-1} . The change in speed is:

$$\Delta v = v - u = 2.0 - 3.0 = -1.0 \text{ m s}^{-1}$$

- b** The velocity has changed from 3.0 m s^{-1} east to 2.0 m s^{-1} west. The change in velocity of the ball is found by vector subtraction, as shown. The change in velocity of the billiard ball is 5.0 m s^{-1} west.



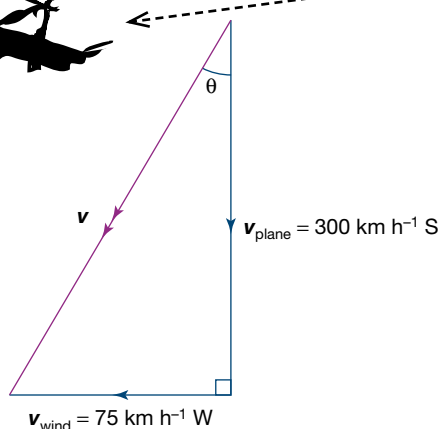
INTERACTIVE TUTORIAL
Relative velocities

- c** To determine the average acceleration of the ball during the collision, the contact time of 25 ms must first be converted to 0.025 s:

$$\text{Acceleration} = \mathbf{a} = \frac{\Delta \mathbf{v}}{\Delta t} = \frac{5.0 \text{ m s}^{-1} \text{ west}}{0.025} = 200 \text{ m s}^{-2} \text{ west}$$

Worked example 1.1C

A pilot is trying to fly her plane due south at an air speed of 300 km h^{-1} . However, it has been blown off course by a crosswind of 75 km h^{-1} to the west. What is the ground speed of the plane, and what is the bearing of its actual flight path?



Solution

This involves a vector addition as shown in the diagram.

The ground speed of the plane is the magnitude of its actual velocity, v , with respect to the Earth. This is the vector sum of its air velocity, v_{plane} , and the wind velocity, v_{wind} :

$$v = v_{\text{plane}} + v_{\text{wind}}$$

Using Pythagoras's theorem: $v = \sqrt{300^2 + 75^2} = \sqrt{95\,625} = 310 \text{ km h}^{-1}$

Using trigonometry to find θ : $\tan \theta = 75/300 = 0.25$, so $\theta = 14^\circ$.

The wind causes the plane to fly at a ground speed of $310 \text{ km h}^{-1} \text{ } 194^\circ \text{T}$.



Figure 1.2 The centre of mass (marked +) of a spanner moves with a constant velocity as it slides across an ice surface, even though the spanner is spinning. Use your ruler to confirm this.

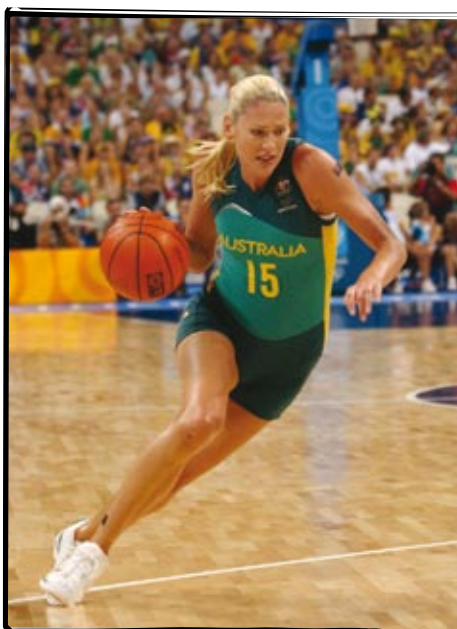


Figure 1.3 Even though champion Australian basketballer Lauren Jackson is running at around 6 m s^{-1} , her actual motion is rather more complicated. Her arm and leg that are swinging forwards are travelling faster than 6 m s^{-1} , while her arm that is swinging back is moving slower than this. Her foot that is in contact with the ground is stationary! To simplify this complex motion, we treat her as a point mass located at her centre of mass.

Centre of mass

The motion of many objects is quite complex. For example, as a person runs, their arms and legs swing in different directions and with different speeds from the rest of their body. To analyse the motion of each of these body parts would be very complicated, but we can simplify this system as a single mass located at a single point. This point is called the *centre of mass* of the system. For most people, the centre of mass is just above the waist. An athlete sprinting down a track at 9 m s^{-1} with arms and legs in rapid motion can thus be treated as a single point mass moving down the track at 9 m s^{-1} .

The significance of the centre of mass is that it enables the motion of a complex object or system to be thought of as a simple point mass located at the centre of mass. Consider the multiflash photograph shown in Figure 1.2. It is the view from above of a spanner sliding and spinning as it travels across an ice surface. The centre of mass of the spanner is marked with a cross.

The positions of the centre of mass form a straight line and are equally spaced, so the centre of mass of the spanner is moving with a constant velocity. The motion is the same as for a simple point mass: the spanner is behaving as though its total mass is located at its centre of mass.

Graphing motion

The motion of a body in a straight line can be represented in graphical form. Graphs are easier to interpret than a table of data, especially for cases of complicated motion. Graphs with any of distance, displacement, speed, velocity or acceleration can be drawn as a function of time.

Position–time graphs

A position–time graph indicates the position of a body relative to an arbitrary origin as a function of time. Figure 1.4 shows the graph for a swimmer completing two 50 m laps of a pool. For motion in a straight line, the direction represented by a positive displacement (e.g. left to right, east to west) must be agreed upon. The *velocity* of the body can be determined from the position–time graph as the *gradient* over the time interval in question. An instantaneous velocity can be found from a curved graph as the gradient of the tangent to the line at the time of interest. This is shown in Worked example 1.1D.

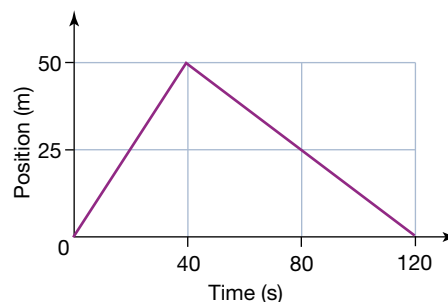


Figure 1.4 A graph of the motion of a swimmer travelling 50 m in a pool, then turning and swimming back to the starting position. The second lap is slower than the first.

Velocity–time graphs

A velocity–time (v – t) graph provides information about the speed and direction of a body during its motion. The sign of the velocity indicates the direction of the motion, but, again, there must be agreement about what, say, a positive direction actually represents. Where the graph cuts the time axis, the body is stationary. The *displacement* of the object being investigated can be found as the *area* under the graph. The total displacement will be the vector sum of all the component displacements (i.e. the sign of the area is included in the calculation), and the total distance covered will be the scalar sum of the areas. Furthermore, the average *acceleration* of the body whose motion is being investigated can be found as the *gradient* of the graph, for the relevant time interval.

Figure 1.5 describes the motion of a dancer moving back and forth across a stage during a 6-second movement. Assume that motion to the right is in the positive direction. The dancer initially moves at 4 m s^{-1} towards the left. For the first 2 s, he continues to move towards the left while slowing down. After 2 s, the dancer stops, then moves towards the right of the stage and speeds up for 2 s. For the final 2 s, he moves with a constant velocity of 4 m s^{-1} towards the right.

The gradient of the graph gives the acceleration of the dancer—a constant acceleration of $+2 \text{ m s}^{-2}$ for the first 4 s. This applies when he is moving left and slowing down, when momentarily stationary, and when moving right and speeding up. For the final 2 s, he is moving with a constant velocity and so has an acceleration of zero.

The graph can also be used to calculate the displacement of the dancer at any time. By finding the area between the line and the time axis, we see that during the first 2 s, he moves 4 m to the left. After 4 s, his displacement is zero (i.e. he is back at his starting position). After 6 s, he has a displacement of +8 m, so he finishes the movement 8 m to the right of his starting position.



PRACTICAL ACTIVITY 2

Locating the centre of mass

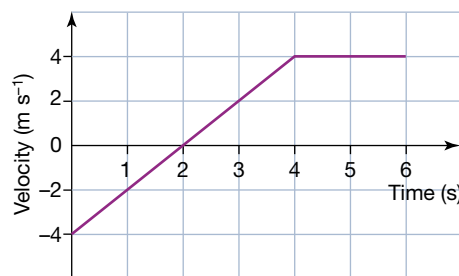
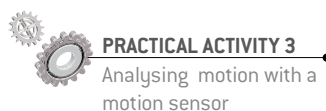


Figure 1.5 This v – t graph represents the motion of a dancer moving back and forth across a stage.

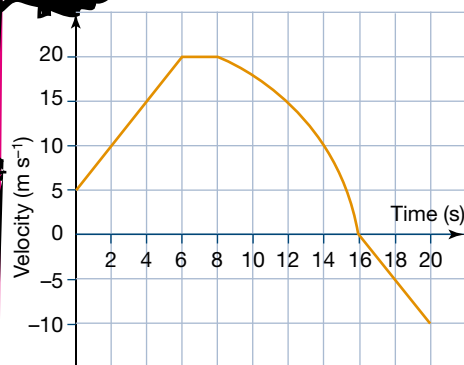


PRACTICAL ACTIVITY 3

Analysing motion with a motion sensor

Acceleration–time graphs

An acceleration–time (a – t) graph indicates the acceleration of the body at any time. The *area* under an a – t graph is the *change in velocity* for the object over the time in question. To establish the actual velocity of the object, the initial velocity must be known.



Worked example 1.10

A car being driven by a learner-driver travels along a straight-line path for 20 s. The graph represents the motion of the car. The positive direction is understood to be east.

- In general terms, describe the motion of the car.
- For how long does the car move in an easterly direction?
- What is the average acceleration of the car during the first 6 s?
- What is the instantaneous acceleration of the car after 12 s?
- What is the displacement of the car during its 20 s journey?
- Calculate the average velocity of the car over the 20 s.

Solution

- The car is initially moving east at 5.0 m s^{-1} . It then accelerates uniformly for 6 s until it reaches a velocity of 20 m s^{-1} east. It maintains this velocity for 2 s but then starts to slow down, still moving east. It stops and turns around at $t = 16 \text{ s}$, accelerating west for the last 4 s.
- The velocity of the car is positive for the first 16 s of its motion, and it is therefore moving east during this time.
- The acceleration during the first 6 s is constant and can be found by calculating the gradient of the line over the first 6 s.

$$\text{Acceleration} = \text{gradient} = \frac{20 - 5}{6 - 0} = +2.5 \text{ m s}^{-2}$$
- To find the acceleration at 12 s, it is necessary to draw a tangent to the curve and calculate its gradient. This gives an instantaneous acceleration of:

$$\text{Acceleration} = \text{gradient} = \frac{\text{rise}}{\text{run}} = -1.9 \text{ m s}^{-2}$$
- The displacement is the area under the v - t graph. 'Counting squares' is a good way of finding this. For this graph, each 'square' represents a displacement of $5 \times 2 = 10 \text{ m}$. The displacement between 8 and 16 s is an estimation owing to the curved nature of the line. The overall displacement for the 20 s is approximately $75 + 40 + 110 - 20 = +205 \text{ m}$, i.e. 205 m towards the east.
- The average velocity over the 20 s time interval will be:

$$v_{\text{av}} = \frac{\text{displacement}}{\text{time}} = \frac{205 \text{ m east}}{20} = 10.3 \text{ m s}^{-1} \text{ east}$$

Physics file

The equations of motion can be derived from the definitions for velocity and acceleration.

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

By rearrangement, $at = \Delta v$, so that:

$$at = v - u \text{ or } v = u + at$$

$$\text{Also, } v_{\text{av}} = \frac{x}{t} \text{ or } x = v_{\text{av}} t$$

Substituting for v_{av} :

$$x = \frac{1}{2}(u + v)t$$

and substituting $v = u + at$ for v :

$$x = \frac{1}{2}(u + u + at)t$$

$$x = ut + \frac{1}{2}at^2$$

Try to derive the other equations yourself.

Equations of motion

When an object is moving in a straight line with constant acceleration, the following equations of motion developed in Year 11 are used to quantitatively describe its motion. The equations are:

$$x = \frac{1}{2}(u + v)t$$

$$v = u + at$$

$$x = ut + \frac{1}{2}at^2$$

$$x = vt - \frac{1}{2}at^2$$

$$v^2 = u^2 + 2ax$$

Notice that each equation contains four of the five quantities relating to the motion of the body. A quick method of deciding which equation to use is to look at which quantity you do not need. For example, if you know u , v and a and need to find x , the only quantity that you are *not* concerned with is t . A quick glance through the five equations reveals that the last one ($v^2 = u^2 + 2ax$) is the only one that does not contain t , so that is the one you should use.

When solving problems using the equations of motion, the following steps are useful.

- 1 Try to visualise what is happening in the problem and draw a simple diagram of the situation.
- 2 If the problem involves a change of direction, call one direction positive and the other negative.
- 3 Identify all the facts given in the question so that you can select an equation that will solve the problem for you.
- 4 Show your working, and make sure that you use the appropriate number of significant figures in your answer.
- 5 Include units with the answer, and specify the direction if the quantity is a vector.

Vertical motion

An object in free-fall near the Earth will accelerate downwards. The Earth's gravity is responsible for this. If air resistance is ignored, this acceleration is uniform and is equal to 9.8 m s^{-2} down. For example, when a bungee jumper falls, their speed will increase by 9.8 m s^{-1} each second until the cord restrains them. When a tennis ball is hit vertically into the air, its speed will decrease by 9.8 m s^{-1} each second until it reaches maximum height. Then it will speed up by 9.8 m s^{-1} each second as it falls back. Since objects in free-fall have a *constant acceleration*, the *equations of motion* can be used to analyse their motion.

Worked example 1.1E

Alicia is bouncing vertically on a trampoline. Her highest bounce was 3.5 m.

- a How long did it take her to reach this height?
- b What was her initial speed?
- c What was her acceleration at the maximum height?
- d What was her velocity 1.5 s after leaving the trampoline?

Solution

Take the direction up as being positive.

- a At the maximum height, Alicia's velocity is zero.

$$x = 3.5 \text{ m}, v = 0, a = -9.8 \text{ m s}^{-2}, t = ?$$

$$x = vt - \frac{1}{2}at^2$$

$$3.5 = 0 - \frac{1}{2}(-9.8)t^2$$

$$t = \sqrt{0.71}$$

$$= 0.85 \text{ s}$$

- b Consider her bounce up to the maximum height.

$$x = 3.5 \text{ m}, v = 0, a = -9.8 \text{ m s}^{-2}, u = ?$$

$$v^2 = u^2 + 2ax$$

$$0 = u^2 + 2 \times -9.8 \times 3.5$$

$$u = \sqrt{68.6}$$

$$= 8.3 \text{ m s}^{-1}$$

Her initial speed is 8.3 m s^{-1} .

- c Her acceleration is always 9.8 m s^{-2} down while she is in mid-air. Her velocity is zero at the maximum height, but her acceleration at this point is 9.8 m s^{-2} down or -9.8 m s^{-2} .
- d Consider her motion from the beginning of the bounce.

$$u = 8.3 \text{ m s}^{-1}, t = 1.5 \text{ s}, a = -9.8 \text{ m s}^{-2}, v = ?$$

$$v = u + at$$

$$= 8.3 + (-9.8 \times 1.5)$$

$$= -6.4 \text{ m s}^{-1}$$

Her velocity is 6.4 m s^{-1} down.

Frames of reference

Imagine that you are in a car that is travelling along a straight, flat stretch of freeway at 100 km h^{-1} . You have an apple in your hand that you toss up and down and side to side. In the car, we can say that your frame of reference is moving with a constant velocity relative to the ground. Now, say that you repeat exactly the same actions on the apple when the car is stationary. The motion of the apple, its mass and acceleration are exactly the same in this stationary reference frame as they were when the car was moving with constant velocity.

A *stationary* frame of reference and a frame of reference with *constant velocity* are called *inertial frames of reference*. Newton's laws of motion are valid in these inertial frames of reference.

Newton also assumed that physical quantities such as mass, time, distance and so on were *absolute* quantities. This means that their values did not change whatever the frame of reference. This would seem to make sense. After all, the mass of an apple and the length of a metre ruler don't change as they travel faster—or do they?

About 200 years after Newton published his laws of motion, Albert Einstein showed that Newton's laws did not work at speeds approaching the speed of light. In fact, at these high speeds, the mass of an object is greater, time slows down and lengths shrink! These ideas are outlined in Einstein's theory of special relativity (see Chapter 6 'Einstein's special relativity'). In this theory, Newton's ideas of the absolute nature of space and time were replaced by Einstein's ideas of the relative nature of space and time. In fact, history has shown that Newton's laws were a special case of Einstein's theories, applying only to situations involving comparatively slow-moving objects.

Relative motion

Is it possible for one car travelling at 100 km h^{-1} to collide with another car travelling at 99 km h^{-1} and there to be no serious damage or injuries? You might have realised that this could well be the case as long as the two cars are travelling in the same direction. The amount of damage that results from a collision depends not so much on how fast the cars

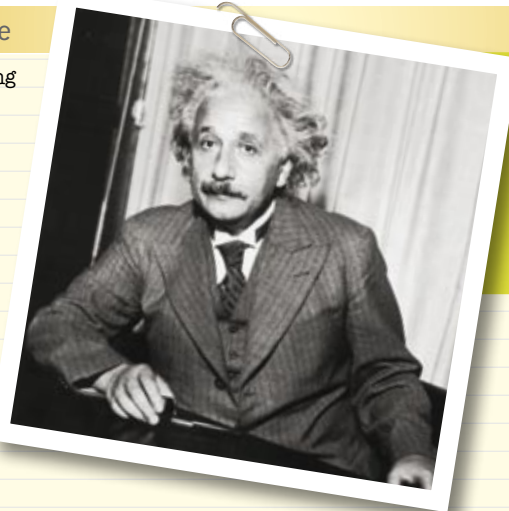


Figure 1.6 Albert Einstein's theory of special relativity was published in 1905 when he was 26 years old. In this theory, Einstein described the relativistic nature of the Universe. These ideas came to replace Newton's theories, which were based on the absolute nature of the Universe.

are moving, but on what their velocity is *relative* to each other. If the two cars had collided while travelling in opposite directions, the consequences would have been catastrophic for the cars and their occupants.

When you discuss the velocity of an object, you usually assume, without saying so, that the frame of reference is the Earth. If you describe an emu as walking with a constant velocity, you really mean that the emu has a constant velocity relative to Earth. In the past, this omission has not been an issue because the frame of reference has usually been Earth. In situations where we are analysing motion from a different frame of reference, the frame needs to be stated.

Imagine that you are in train A that is moving east with a constant speed of 10 m s^{-1} and you are walking along the aisle towards the front of the train at 2 m s^{-1} . What velocity are you travelling at? The answer to this question is that it depends on which frame of reference you are using. To a person sitting in the train, your velocity is 2 m s^{-1} east. Here, your frame of reference is the moving train. However, to a person standing on the station platform, your velocity will be 12 m s^{-1} east. In this case, your frame of reference is the ground or Earth. Now imagine that you sit down and that another train, train B, passes in the opposite direction at 5 m s^{-1} relative to the ground. What is the velocity of this train relative to your train?

The velocity of train B relative to train A gives the apparent motion of train B when seen from train A. In other words, this gives train A's view of how fast train B seems to be moving.

From your frame of reference in train A, train B seems to be travelling to the west but faster than 5 m s^{-1} . To calculate the relative velocity of two objects, we need to perform a *vector subtraction*.

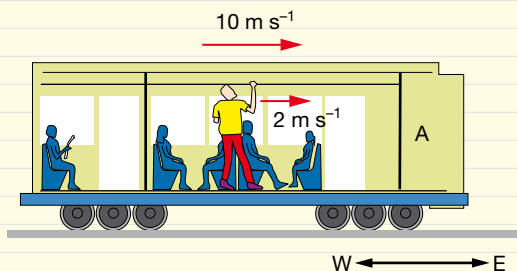


Figure 1.7 The velocity of this person relative to the ground is 12 m s^{-1} east, but their velocity relative to the other passengers is 2 m s^{-1} east.



Velocity of B relative to A = velocity of B relative to ground – velocity of A relative to ground:

$$v_{B,A} = v_{B,G} - v_{A,G}$$

As shown in Figure 1.8, this gives a relative velocity of 15 m s^{-1} west.

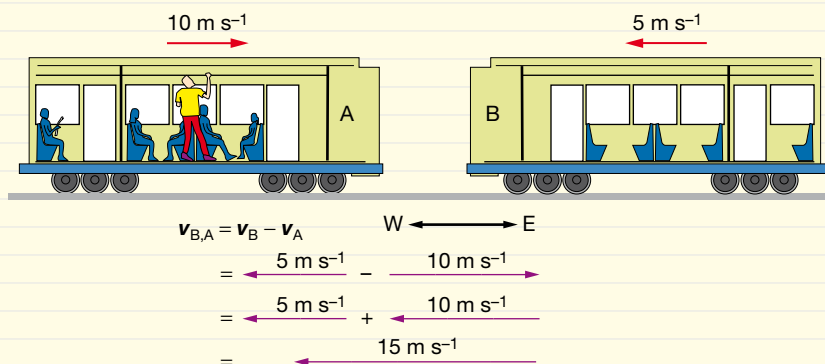


Figure 1.8 To determine the velocity of train B relative to train A, a vector subtraction is performed, giving a relative velocity of 15 m s^{-1} west.



1.1 summary

Mechanics review

- Physical quantities can be divided into scalars and vectors; scalars can be described fully by a magnitude, while vectors must have both magnitude and direction.
- Vectors can be added and subtracted.
- An object's centre of mass is the point at which the total mass of the object can be considered to be concentrated.
- A position–time graph indicates the position of a body relative to an arbitrary origin, as a function of time.
- A velocity–time (v – t) graph describes the speed and direction of a body during its motion.
- An acceleration–time (a – t) graph indicates the acceleration of a body at any time.
- Five equations are used to quantitatively describe motion with a uniform acceleration:

$$\begin{aligned}
 x &= \frac{1}{2}(u + v)t \\
 v &= u + at \\
 x &= ut + \frac{1}{2}at^2 \\
 x &= vt - \frac{1}{2}at^2 \\
 v^2 &= u^2 + 2ax
 \end{aligned}$$
- If air resistance can be ignored, an object in free-fall will have a constant acceleration of 9.8 m s^{-2} down.



1.1 questions

Mechanics review

In the following questions, assume that the acceleration due to gravity is 9.80 m s^{-2} down.

- Label S for scalar and V for vector in the following list:

distance___ displacement___ time___
 velocity___ speed___ acceleration___
 momentum___ energy___

- Which of the following is represented by the area under a velocity–time graph?
 - Average acceleration over the time interval
 - Average velocity over the time interval
 - The displacement during the time interval
 - The change in velocity for the object over the time interval



- 3 The following velocity–time graph was derived from a test drive of a prototype sports car. The car started from rest and initially travelled north.



- a What distance did the car travel:
 - i during the first 4 seconds of its motion?
 - ii between 4 s and 12 s?
 - iii between 12 s and 28 s?
 - b Calculate the displacement of the car after 28 s.
 - c What was the total distance that the car travelled during the trip?
 - d What was the average speed of the car during the 28 s interval?
 - e Calculate the acceleration of the car between 4 s and 12 s.
- 4 A car travelling with a constant speed of 80 km h^{-1} passes a stationary motorcycle policeman. The policeman sets off in pursuit, accelerating uniformly to 80 km h^{-1} in 10.0 s and reaching a constant speed of 100 km h^{-1} after a further 5.0 s. At what time will the policeman catch up with the car?
- In questions 5–10, ignore the effects of air resistance.
- 5 A golf ball is dropped from the top of a sheer cliff 78.0 m above the sea.
 - a Calculate the velocity of the golf ball after 1.00 s.
 - b How far will the ball fall in the first 2.00 s?
 - c Calculate the speed of the golf ball after it has travelled 10.0 m.
 - d When will the golf ball hit the water?
 - e What is the acceleration of the golf ball at:
 - i 1.50 s?
 - ii 3.50 s?
 - 6 Andrew is dragging a 60 kg crate of books across the garage floor. He pulls with a force of 120 N on a rope that is at an angle of 35° to the horizontal.
 - a Calculate the horizontal component of the pulling force.
 - b Calculate the vertical component of the pulling force.
 - 7 Cassie starts from rest at the top of a 3.2 m long playground slide and slides to the bottom with a constant acceleration. If she takes 2.4 s to reach the bottom, calculate:
 - a her average speed
 - b her average acceleration
 - c her final speed
 - d her speed when she is halfway down the slide.
 - 8 Vinh is investigating the bouncing ability of a golf ball and a tomato. He drops both objects from a height of 2.00 m and measures the rebound heights. He found that the golf ball rebounded to 1.50 m and the tomato just splattered without rebounding at all.
 - a What was the speed of the golf ball just before it landed?
 - b What was the speed of the tomato just before it hit the ground?
 - c Calculate the speed of the golf ball as it rebounded.
 - d Which object had the greater change in speed as it landed? Calculate the speed change of this object.
 - e Which of these objects experienced the greater change in velocity as it landed? Calculate the velocity change of this object.
 - 9 Fiona rides her skateboard up a ramp. She begins with a speed of 8.0 m s^{-1} but slows with a constant deceleration of 2.0 m s^{-2} . She travels some distance up the ramp before coming to rest, then rolls down again. Ignoring air resistance and friction, calculate:
 - a the distance that Fiona travels up the ramp before stopping
 - b the time that it takes Fiona to reach this highest point
 - c Fiona's velocity after 5.0 s has elapsed.
 - 10 A plane is flying due south at 100 m s^{-1} on an initially still day. Then a crosswind with a velocity of 25.0 m s^{-1} towards the west begins to blow.
 - a What is the velocity of the plane under the influence of this crosswind?
 - b In which direction should the pilot steer the plane to maintain a velocity of 100 m s^{-1} south?

1.2 Newton's laws of motion

When Newton published his three laws of motion in 1687, he revolutionised our understanding of the physical world. Up until then, people thought differently about why things moved the way they did. The ideas of the ancient Greeks were prevalent. These ideas, proposed by Aristotle, were that there were two types of motion—natural and violent. It was also thought that when a constant force acted on an object, the object would move with a constant speed. To many people even today, these ideas seem to be correct. In fact, Galileo and Newton showed them to be incorrect over 300 years ago!

Today, even though Einstein's theories have superseded those of Newton, we still use Newton's laws for most situations. After a car accident, investigators will use Newton's laws to analyse the motion of the vehicles. When NASA scientists program the courses of spacecraft on flights to Mars, Jupiter and beyond, they use Newton's laws. Solar and lunar eclipses can be predicted with great precision many centuries into the future by using Newton's laws. Similarly, they can be applied to predict times and sizes of the tides that alter the ocean depths around the globe. In fact, it is only in situations involving extremely high speeds (greater than 10% of the speed of light) or strong gravitational fields that Newton's laws become imprecise and Einstein's theories must be used.

Newton's laws describe how the concept of force can be used to explain why a body moves in the way that it does. The first law describes what happens to a body when it experiences zero net force (i.e. $\Sigma F = 0$). Galileo had previously called this behaviour *inertia*. The second law explains how the body will respond when an unbalanced force is acting (i.e. $\Sigma F \neq 0$), and the third law states that all forces act in pairs, known as *action–reaction pairs*.



NEWTON'S FIRST LAW states that every object continues to be at rest, or continues with constant velocity, unless it experiences an unbalanced force. This is also called the law of inertia.



NEWTON'S SECOND LAW states that the acceleration of a body experiencing an unbalanced force is directly proportional to the net force and inversely proportional to the mass of the body:

$$\Sigma F = ma$$

The net (or resultant) force, ΣF , is measured in newtons (N), when the mass is measured in kilograms (kg) and the acceleration, a , is measured in metres per second squared (m s^{-2}).



NEWTON'S THIRD LAW states that when one body exerts a force on another body (an action force), the second body exerts an equal force in the opposite direction on the first (the reaction force):

$$F(A \text{ on } B) = -F(B \text{ on } A)$$

An important aspect of Newton's third law is that the action and reaction forces act on *different* bodies. This means that action/reaction pairs will **never be added together**.

If a young child accidentally runs into a wall, the action force will be applied to the wall, but the reaction force will be applied to the child in the opposite direction. These forces will *always be equal* in magnitude, but the



PRACTICAL ACTIVITY 4

Newton's laws of motion



PRACTICAL ACTIVITY 5

Action and reaction



Figure 1.9 Sir Isaac Newton (1642–1727) is widely considered to be the greatest scientist of all. He developed the scientific method of experimental research. This paved the way for others to extend the boundaries of knowledge and led to the scientific and technological revolution that has transformed the world over the past 300 years.



Figure 1.10 The passengers in this aeroplane are moving with a constant velocity. The forces acting on them, according to Newton's first law, are in balance.

Physics file

Newton's book *Principia Mathematica* is one of the most important publications in the history of science. At the start of this book, Newton wrote down his assumptions about the absolute nature of space and time. He wrote 'The following two statements are assumed to be evident and true. Absolute, true and mathematical time, of itself, and from its own nature, flows equably without relation to anything external. Absolute space, in its own nature, without relation to anything external, remains always similar and immovable.'



PRACTICAL ACTIVITY 6

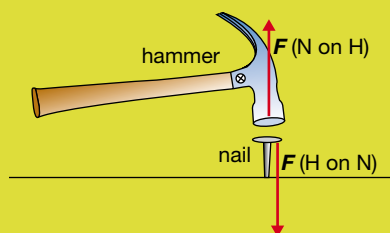
Newton's second law

effect of the forces will be different. Since the masses involved are so different, the resulting accelerations of the wall and the child's body will also be very different! The reaction force will cause the child to experience a rather large deceleration, but the same-sized action force acting on the much more massive wall will not change its motion to any measurable degree.



Figure 1.11 The passengers in this roller-coaster ride are accelerating. The forces acting on them, according to Newton's second law, are unbalanced.

(a)



(b)



(c)

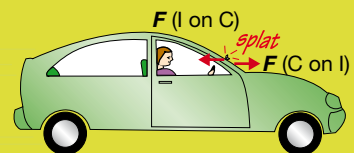


Figure 1.12 Action/reaction forces.

- (a) The hammer exerts a downward force on the nail, and the nail exerts an upward force on the hammer; these forces are equal in magnitude.
- (b) The rocket exerts a downward force on the fuel and gases, and the fuel and gases exert an upward force on the rocket; these forces are equal in magnitude.
- (c) The car exerts a forward force on the insect, and the insect exerts a backward force on the car; these forces are equal in magnitude.

Worked example 1.2A

Santo drags a 60 kg Christmas tree across a floor at a constant speed of 1.5 m s^{-1} . If the force of friction between the tree and the floor is 50 N and it is being pulled at an angle of 35° to the horizontal, calculate:

- a** the net force on the tree
- b** the force that Santo exerts on the tree
- c** the force that the tree exerts on Santo.

Solution

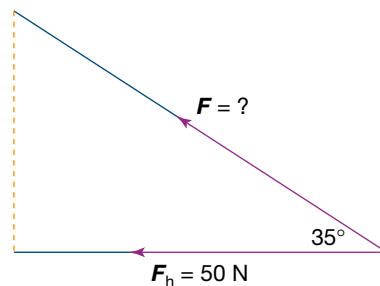
- a** The tree has a constant velocity so, from Newton's first law, there is zero net force acting on it.
- b** If the net force on the tree is zero, the horizontal forces must be in balance. This means that the frictional force of 50 N must be equal but opposite to the horizontal component of the pulling force.

$$F_h = F \cos 35^\circ$$

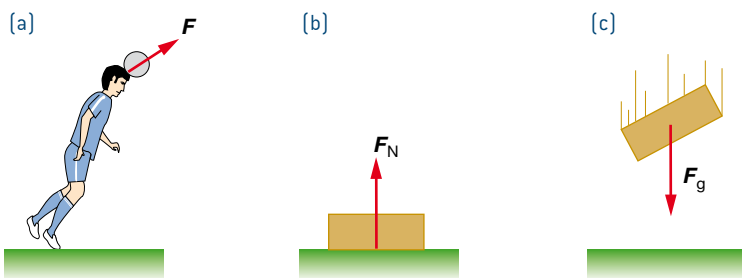
$$50 = F \times 0.82$$

So, $F = 61$ N [i.e. the tree is being pulled with a force of 61 N at an angle of 35° above the horizontal].

- c** The force that Santo exerts on the tree and the force that the tree exerts on Santo are an action/reaction pair. According to Newton's third law, they are equal and opposite. Thus the tree must exert a force on Santo of 61 N at an angle of 35° below the horizontal.



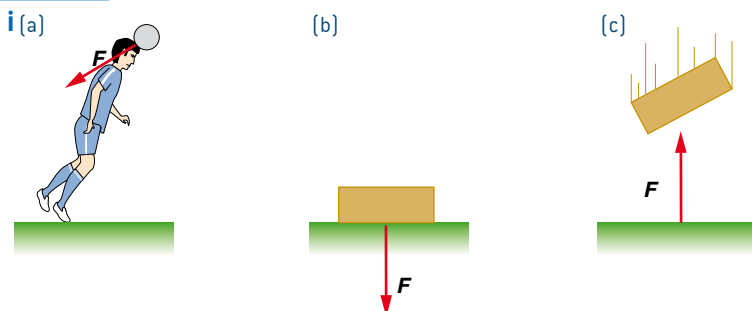
Worked example 1.2B



In each of the situations shown, one of the forces in the action/reaction pair is given.

- In each case, draw the other force as suggested by Newton's third law, being mindful of its location and size.
- For each of the forces that you have drawn, complete the following statement: Force exerted on _____ by _____.
- The soccer player in part (a) stated that since the force exerted by the ball on his head was equal in magnitude to the force exerted by his head on the ball, then the forces should cancel out and the ball should stay where it is. This argument is obviously flawed. Where is the mistake in his reasoning?

Solution



- Force exerted on head by ball.
 - Force exerted on floor by brick. [Note: This is not the weight force F_g .]
 - Gravitational force of attraction exerted on Earth by brick.
- It is incorrect to say that the forces cancel out. One of the forces is acting on his head; the other is acting on the ball. Forces can only be added if they are acting on the same object.

Physics file

In this book, we will use the subscript notation for forces. Gravity is denoted by F_g . W is also commonly used for gravity. The normal force is given by F_N . N is also commonly used for this force.

Other forces that you will encounter include the force from a spring, F_s , friction, F_f , air resistance, F_a , and the force of tension, F_t .

Air resistance, speed skydiving and skysurfing

Air resistance (or drag) acts whenever a solid or a liquid has to travel through air. For example, when a raindrop falls, it collides with millions of air molecules which impede its motion. In fact, if it were not for air resistance, raindrops would be travelling at thousands of kilometres per hour by the time they reached the ground! Air resistance is a force that particularly affects objects with small masses and large surface areas. Hang-gliders are light and have a large wing area, so their motion is strongly influenced by the air. Air resistance increases with the speed of an object and is also greatly affected by the object's shape. Streamlined bodies pass through the air more easily.

In many activities such as cycling, skiing and car racing, the objective is to reduce the amount of air resistance in order to travel as fast as possible. In speed skydiving (see Figure 1.13a), the divers jump from an altitude of 4000 m and then their average speed between 2700 m and 1700 m is recorded. The world record in 2007 was 504 km h⁻¹. The divers adopt a very streamlined position as they fall in order to minimise the drag forces acting on them.

Air resistance (F_a) varies as the square of the speed (v) of an object. Mathematically, this can be expressed as:

$$F_a \propto v^2$$

$$\text{or } F_a = kv^2$$

The constant k depends on the density (ρ) of the air, the surface area (A) perpendicular to the motion and the drag coefficient (C_d) of the object. The drag coefficient indicates how smoothly the air flows past the object. Spheres have a drag coefficient of about 0.5, but it can be as high as 2 for oddly shaped bodies. The complete relationship for air resistance is:

$$F_a = \frac{1}{2}\rho AC_d v^2$$

In skysurfing, the objective is to increase the amount of air resistance in order to fall as slowly as possible and extend the amount of time spent in the air. Skysurfers attach their feet to a small surfboard, place a parachute on their back and then jump from aeroplanes from an altitude of several kilometres. As they fall, skysurfers can reach speeds of around 170 km h⁻¹. If the board is horizontal, the skysurfer will fall vertically. However, by tipping the board forwards, backwards or sideways, skysurfers are able to control their motion to some extent and 'surf' through the air as they fall towards the ground. Extending their arms also increases the amount of drag that is acting.

To estimate the drag coefficient of the skysurfer, we can assume that the skysurfer and the equipment have a total mass of 75 kg and that they are falling vertically with a constant velocity of 170 km h⁻¹ on a board with an area of about 0.20 m². If the arms of the skysurfer are extended, they contribute an additional area of about 0.20 m². The density of air is 1.25 kg m⁻³. The only forces acting on the skysurfer during the fall are the gravitational force, F_g or W , and air resistance, F_a . The skysurfer is falling with a constant velocity, and so these forces are in balance, i.e. $F_a = F_g$, so:

$$F_a = F_g = mg$$

$$= 75 \times 9.8$$

$$= 735 \text{ N}$$

$$\rho = 1.25 \text{ kg m}^{-3}, A = 0.40 \text{ m}^2, v = 170 \text{ km h}^{-1} = 47 \text{ m s}^{-1}$$

$$\text{Since } F_a = \frac{1}{2}\rho AC_d v^2$$

$$735 = 0.5 \times 1.25 \times 0.40 \times C_d \times 47^2$$

$$\text{So } C_d = 1.33$$

This would seem to be a reasonable value for the drag coefficient, given the irregular shape of the skysurfer.



Figure 1.13 A major consideration in many competitive sports is the effect of air resistance. (a) Greater speed can be attained by this speed skydiver if the effects of air resistance are minimised by streamlining. (b) Skysurfers want to extend the time it takes to fall and need to maximise the amount of drag. They make use of the air resistance to manoeuvre the board and to perform stunts.



1.2 summary

Newton's laws of motion

- Newton's first law states that a body will either remain at rest or continue with constant velocity unless it is acted upon by an unbalanced force.
- Newton's second law states that the acceleration of a body is directly proportional to the net force acting on it and inversely proportional to its mass. This is usually written as:

$$\Sigma F = ma$$
- Newton's third law states that when one body exerts a force on another body (an action force), the second body exerts an equal force in the opposite direction on the first (the reaction force):

$$F(\text{A on B}) = -F(\text{B on A})$$
- It is important to remember that the action/reaction forces act on *different* objects. It is therefore incorrect to add them and so they never cancel each other.



1.2 questions

Newton's laws of motion

- Vickie's car has a flat battery and will not start. She gets a few friends to help push her car so that she can jump-start it. Vickie is surprised to find that it takes six friends pushing to start the car rolling, but only one friend pushing to keep the car rolling. Explain, in terms of Newton's laws, why this is so.
- Phil is standing inside a tram when it starts off suddenly. Len, who was sitting down, commented that Phil was thrown backwards as the tram took off. Is this a correct statement? Explain in terms of Newton's laws.
- A table-tennis ball of mass 10 g is falling towards the ground with a constant speed of 8.2 m s^{-1} . Calculate the magnitude and direction of the air resistance force acting on the ball.
- A wayward willy wagtail crashes into a large window and knocks itself out. Which one of the following statements correctly describes the forces that act during this collision?
 - The force that acts on the willy wagtail is greater than the force that acts on the window.
 - The force that acts on the willy wagtail is smaller than the force that acts on the window.
 - The force that acts on the willy wagtail is the same size as the force that acts on the window.
 - The forces that act on the window and willy wagtail cancel each other out.
- Ishtar is riding a motorised scooter along a level bike path. The combined mass of Ishtar and her scooter is 80 kg. The frictional and drag forces that are acting total to 45 N. What is the magnitude of the driving force being provided by the motor if:
 - she is moving with a constant speed of 10 m s^{-1}
 - she is accelerating at 1.5 m s^{-2}
- A cyclist and his bike have a combined mass of 80 kg. When starting off from traffic lights, the cyclist accelerates uniformly and reaches a speed of 7.5 m s^{-1} in 5.0 s.
 - What is the acceleration during this time?
 - Calculate the driving force being provided by the cyclist's legs as he starts off. Assume that frictional forces are negligible during this time.
 - The cyclist keeps riding along with a constant speed of 7.5 m s^{-1} . Assuming that the force being provided by his legs is now 60 N, determine the magnitude of the frictional forces that are acting.
- During preseason football training, Matt was required to run with a bag of sand dragging behind him. The bag of mass 50 kg was attached to a rope, which made an angle of 25° to the horizontal. When Matt ran with a constant speed of 4.0 m s^{-1} , a frictional force of 60 N was acting on the bag.
 - What was the net force acting on the bag of sand?
 - Calculate the size of the tension force acting in the rope.
 - What was the magnitude of the force that the rope exerted on Matt as he ran?
- During a game of table tennis, the ball is hit by the bat. Due to this contact, there are forces acting on both the ball and the bat. How do these forces compare in:
 - magnitude?
 - direction?
- A fully laden supermarket trolley is stationary at a check-out. The shopper wishes to push the trolley to their car. When they push the trolley, it will, according to Newton's third law, exert an equal but opposite force on them. Will the trolley move? Explain.
- Complete each of these force diagrams, showing the reaction pair to the action force that is shown. For each force that you draw, state what the force is acting on and what is providing the force.
 -
 -
 -
 -

1.3

The normal force and inclined planes

In section 1.2, we reviewed Newton's laws of motion in one dimension. We saw that when unbalanced forces act on an object, it will accelerate. In this section, we will consider some examples of motion in two dimensions. Once again, Newton's laws will be used to analyse these situations.

Normal forces

If you exert a force against a wall, Newton's third law says that the wall will exert an equal but opposite force on you. If you push with greater force, as shown in Figure 1.14, the wall will also exert a greater force. The force from the wall acts at right angles to the surface, i.e. it is *normal* to the surface and is thus called a *normal force*. Like every force, a normal force is one half of an action/reaction pair, so it is often called a *normal reaction force*. In this book, we will use F_N for the normal force although N is also commonly used.



A **NORMAL** force, F_N or N , acts at right angles to a surface.

During many interactions and collisions, the size of the normal force *changes*. For example, when a ball bounces, the forces that act on it during its contact with the floor are its weight, F_g or W , and the normal force, F_N , from the floor. As can be seen in Figure 1.15, the normal force is not constant, but changes in magnitude throughout the bounce. When contact has just been made, the ball is compressed only slightly, indicating that the force from the floor is minimal. This force then becomes larger and larger, causing the ball to become more and more deformed. At the point of maximum compression, the normal force is at its maximum value. The normal force from the floor is greatest when the bouncing ball is stationary.

The forces acting on a ball as it bounces (its weight, F_g , and the normal force, F_N) are *not* an action/reaction pair. Both *act on the same body*, whereas Newton's third law describes forces that bodies exert on each other. A pair of action/reaction forces that act during the bounce are the upward force, F_N , that the floor exerts on the ball and the downward force that the ball exerts on the floor. This downward force is equal in magnitude to the normal force, so it varies during the bounce.

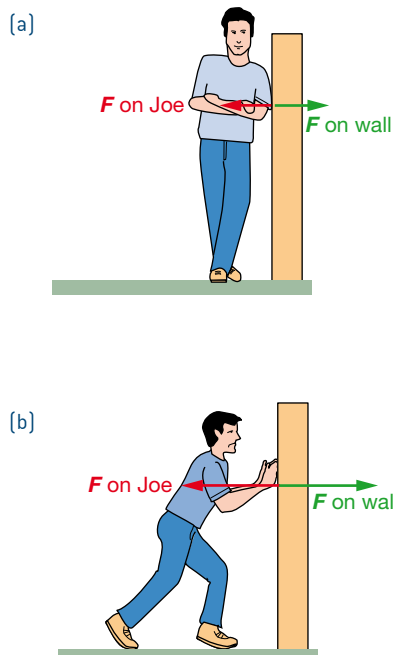


Figure 1.14 (a) If Joe exerts a small force on the wall, the wall will exert a small force on Joe. (b) When Joe pushes hard against the wall, it pushes back just as hard!

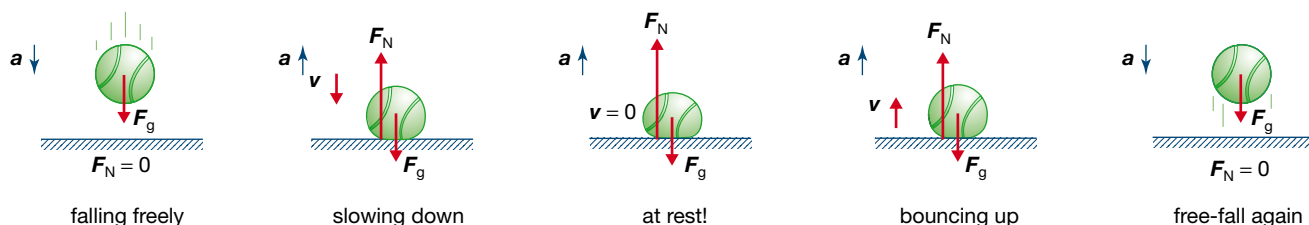


Figure 1.15 The forces acting on a bouncing ball.

Inclined planes

If a toy car is placed on a sloping surface, it will *accelerate* uniformly down the slope. If the angle of the slope is increased, the car will accelerate at a greater rate. To understand this motion, we need to examine the motion of the toy car as it rolls freely across a smooth horizontal surface. The forces acting on the car are its weight, F_g or W , and the normal force, F_N or N , from the surface. If friction is ignored, the car will move with a constant horizontal velocity. The

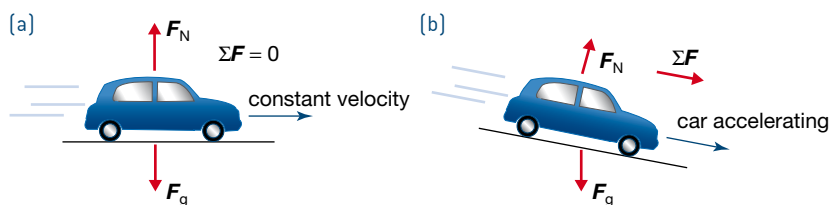


Figure 1.16 The same two forces, F_g and F_N , act on the car when it is (a) rolling horizontally or (b) rolling down an incline. However, these forces are not in balance when the car rolls downhill and so the car accelerates.

car has no motion in the vertical direction, so (as indicated by Newton's first law) the vertical forces must be in balance; that is, $F_N = -F_g$ (Figure 1.16a).

However, if the car now rolls down a smooth plane inclined at an angle θ to the horizontal, the car will *accelerate*. The forces acting on the car are *unbalanced*. If friction is ignored, the only forces acting on the car are still its weight, F_g , and the normal force, F_N , but these forces cannot be balanced because they are not opposite in direction (Figure 1.16b). When the forces are added, they give a net force, ΣF , that is directed down the incline, so the car will accelerate in that direction.



PRACTICAL ACTIVITY 7

Motion on an inclined plane

The usual method of analysing the forces in this situation is to consider the weight, F_g , as having a component that is *parallel* to the incline and a component that is *perpendicular* to the incline (Figure 1.17a). The car is rolling down the incline, so the force parallel to the incline must be responsible for the car's motion. The parallel component of the weight force has a magnitude of $F_g \sin \theta$. Since the car has no motion in the direction perpendicular to the incline, the normal force, F_N , must be equal in magnitude to the perpendicular component of the weight force. This perpendicular component has a magnitude of $F_g \cos \theta$.

If friction is ignored, the parallel component of the weight down the incline is the net force, ΣF , that causes the car to accelerate down the incline. The acceleration, a , of the car down the slope can then be determined from Newton's second law:

$$\Sigma F = F_g \sin \theta$$

$$\text{so: } ma = mg \sin \theta$$

$$\text{and: } a = g \sin \theta$$

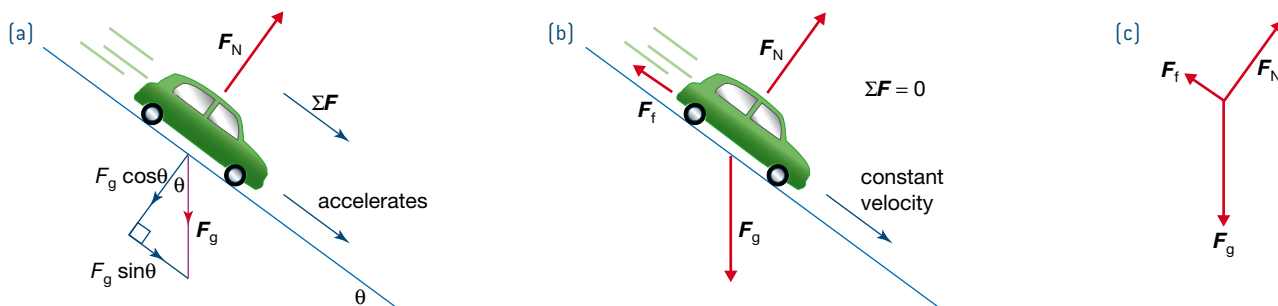


Figure 1.17 (a) The motion of an object on a smooth inclined plane can be analysed by finding the components of the weight force that are parallel and perpendicular to the plane. (b) If friction acts on the car, causing it to move with a constant velocity, the net force on the car will be zero, and so the forces will be in balance both perpendicular and parallel to the incline. (c) A conventional physics diagram of (b) shows the forces acting through the centre of mass.



Acceleration along a smooth incline is given by:
 $a = g \sin \theta$

Physics file

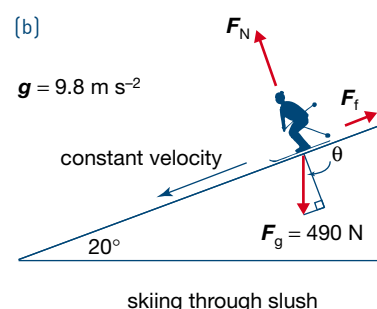
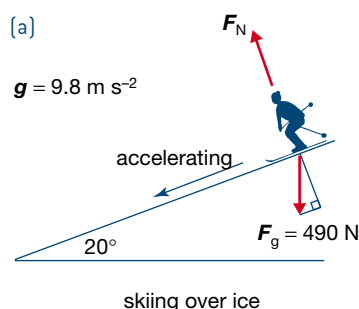
The steepest road in the world is Baldwin St in Dunedin, New Zealand. It has an incline of around 20° at its steepest. This does not sound overly impressive, but angles are deceptive and it is a seriously steep street. In 2000, two university students jumped into a wheelie bin and rolled down Baldwin St. Unfortunately, they crashed into a parked car and suffered serious injuries.

What if one of the wheels of the car becomes stuck, so that now a frictional force, F_f , will act? This will be in a direction opposite to the velocity of the car, i.e. up the incline (Figure 1.17b). When the car is moving down the incline with a constant velocity, the forces acting on the car must be in balance. The forces acting parallel to the incline (the frictional force, F_f , and the parallel component of the weight, $F_g \sin \theta$) will therefore be equal in magnitude; that is, $F_f = F_g \sin \theta$.

Worked example 1.3A

A skier of mass 50 kg is skiing down an icy slope that is inclined at 20° to the horizontal. Assume that friction is negligible and that the acceleration due to gravity is 9.8 m s^{-2} .

- a** Determine:
- the component of the weight of the skier perpendicular to the slope
 - the component of the weight of the skier parallel to the slope
 - the normal force that acts on the skier
 - the acceleration of the skier down the slope.
- b** The skier, while on the same slope, runs into a patch of slushy snow that causes her to move with a constant velocity of 2.0 m s^{-1} . Calculate the magnitude of the frictional forces that act here.



Solution

The diagrams show the forces acting on a skier who is (a) travelling down an icy slope with negligible friction, and (b) skiing through slushy snow with a constant velocity.

- a**
- The weight of the skier $F_g = mg = 50 \times 9.8 = 490 \text{ N}$ down. The perpendicular component of the weight is:

$$F_g \cos \theta = 490 \cos 20^\circ$$

$$= 460 \text{ N}$$
 - The parallel component of the weight is:

$$F_g \sin \theta = 490 \sin 20^\circ = 170 \text{ N}$$
 - The normal force is equal to the component of the weight that is perpendicular to the incline, i.e. $F_N = 460 \text{ N}$.
 - The acceleration of the skier can be determined by considering the forces parallel to the incline. If friction is ignored, the net force acting in this direction is the parallel component of the weight, i.e. 170 N .

$$a = \frac{\Sigma F}{m}$$

$$= \frac{170}{50}$$

$$= 3.4 \text{ m s}^{-2} \text{ down the incline}$$
- b** If the skier is moving with a constant velocity, the forces acting parallel to the slope must be in balance. In other words, the frictional force is equal in magnitude to the parallel component of the weight force (diagram b), i.e. $F_f = 170 \text{ N}$.

The luge and the coefficient of friction

Luge is the French word for sled. It has been a Winter Olympic event since 1964 and the sliders can reach very high speeds. To do this, they need to minimise drag and frictional forces as they race down the course.

When a body slides over a rough surface, friction opposes its motion. The size of the frictional force depends on the nature of the surfaces and the size of the normal force. On a horizontal surface, the normal force is equal to the weight of the body. However, on an inclined plane, the normal force will be less than the weight of the body, and will decrease as the angle of inclination increases. The relationship between the force of friction, F_f , and the normal force, F_N , is expressed as:

$$F_f = \mu F_N$$

The term μ is the coefficient of friction. Its value depends on the type of surfaces involved and whether they are rough, polished, wet, dry and so on. A low coefficient of friction indicates a small degree of friction.

Frictional forces can be measured in different situations, leading to kinetic friction and static friction. *Kinetic friction* (also known as sliding friction) applies when one body is moving across a surface. The coefficient of kinetic friction (μ_k) is used to determine values of kinetic friction. *Static friction* is the force that keeps a body stationary even when a pushing or pulling force is acting on the body. The coefficient of static friction (μ_{st}) is used to calculate values of static friction. Some approximate values of kinetic and static friction for different surfaces are given in Table 1.1. Notice that the *static friction coefficients are always greater than the kinetic friction coefficients*. This is because the forces that act between the surfaces are weakened when the surfaces are moving past each other, and the bonds are not able to form as strongly.

Table 1.1 Coefficients of static friction (μ_{st}) and kinetic friction (μ_k) for various combinations of surfaces

Surfaces	μ_{st}	μ_k
Wood on wood	0.4	0.2
Steel on steel	0.7	0.6
Ice on ice	0.1	0.03
Steel on ice	0.1	0.03
Rubber on dry concrete	1.0	0.8
Rubber on wet concrete	0.7	0.5
Human joints	0.01	0.01
Teflon on Teflon	0.04	0.04

The luge sled sits on a pair of sharp steel runners. At the start of the race, the slider has to paddle to build up speed as quickly as possible. They use gloves with spikes on them that increase the size of the static friction and so grip the ice. They then adopt an aerodynamic position and reduce drag by lying flat on the sled. The sliders steer by applying pressure with their feet to the front runners or kufens of the sled. There are no brakes on the luge! It can be seen in Table 1.1 that the coefficient of kinetic friction for steel on ice is just 0.03. This allows the sliders to reach speeds of around 140 km h^{-1} .



Figure 1.18 Competitors in the luge are known as sliders. They travel feet first down an ice track on a sled without brakes. The luge is so fast that it is timed to one-thousandth of a second.



1.3 summary

The normal force and inclined planes

- A normal force, F_N or N , is the force that a surface exerts on an object that is in contact with it. It acts at right angles to a surface and changes as the force exerted on the surface changes.
- When a ball bounces, the normal force changes throughout the time of contact between the ball and the surface. The normal force is greater than the weight of the bouncing ball for most of the time during their contact, causing the ball to return to the air.



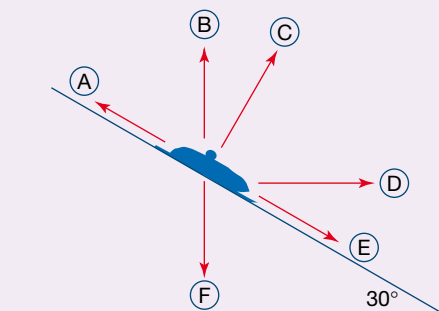
1.3 questions

The normal force and inclined planes

For the following questions, assume that the acceleration due to gravity is 9.8 m s^{-2} and ignore the effects of air resistance.

The following information applies to questions 1–5.

Kirsty is riding in a bobsled that is sliding down a snow-covered hill with a slope at 30° to the horizontal. The total mass of the sled and Kirsty is 100 kg. Initially the brakes are on and the sled moves down the hill with a constant velocity.



- 1 Which one of the arrows (A–F) best represents the direction of the frictional force acting on the sled?
- 2 Which one of the arrows (A–F) best represents the direction of the normal force acting on the sled?
- 3 Calculate the net frictional force acting on the sled.
- 4 Kirsty then releases the brakes and the sled accelerates. What is the magnitude of her initial acceleration?
- 5 Finally, Kirsty rides the bobsled down the same slope but with the brakes off, so friction can be ignored. It now has an extra passenger so that its total mass is now 1400 kg. How will this affect the acceleration of the bobsled?

The following information applies to questions 6–8.

- The normal force, F_N , acting on an object on an inclined plane is equal to the component of the weight force perpendicular to the incline. The steeper the incline, the smaller the normal force.
- For an object that is stationary on a rough inclined plane, the frictional force acts up the incline and is equal in magnitude to the component of the weight force down the slope.
- When an object slides on a smooth inclined plane, its acceleration depends on the angle of the slope:

$$a = g \sin \theta$$

Marshall has a mass of 57 kg and he is riding his 3.0 kg skateboard down a 5.0 m long ramp that is inclined at an angle of 65° to the horizontal. Ignore friction when answering questions 6 and 7.

- 6 a Calculate the magnitude of the normal force acting on Marshall and his skateboard.
b What is the net force acting on Marshall and his board?
c What is the acceleration of Marshall as he travels down the ramp?
- 7 a If the initial speed of Marshall is zero at the top of the ramp, calculate his final speed as he reaches the bottom of the ramp.
b Draw a vector diagram of the forces acting on the skateboarder as he moves down the ramp.
- 8 Marshall now stands halfway up the incline while holding his board in his hands. Calculate the frictional force acting on Marshall while he is standing stationary on the slope.
- 9 A child rolls a 50 g marble up a playground slide that is inclined at 15° to the horizontal. The slide is 3.5 m long and the marble is launched with a speed of 4.8 m s^{-1} .
a What is the magnitude of the normal force acting on the marble as it rolls up the slide?
b How fast is the marble travelling when it is halfway up the slide?
- 10 A tennis ball bounces on a concrete floor. Discuss the relative sizes of the weight of the ball, F_g , and the normal force, F_N , that the floor exerts on the ball when:
a the ball is in mid-air falling towards the floor
b the ball has just made contact with the floor and is slowing down
c the ball is at the point of its maximum compression.

1.4 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 **weight**, which is the force due to gravity, F_g or W . 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, F_g , it follows that the projectile must have a vertical acceleration of 9.8 m s^{-2} downwards.
- The only force, 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, 9.8 m s^{-2} downward.

In the **HORIZONTAL COMPONENT**, a projectile moves with uniform velocity since there are no forces acting in this direction.



Figure 1.19 A multiflash photograph of a tennis ball that has been bounced on a hard surface. The ball moves in a parabolic path.

These points are fundamental to an understanding of projectile motion, and can be seen by studying Figure 1.20.

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 1.21, 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° above the horizontal will land at 50 m s^{-1} at 80° below the horizontal.



PRACTICAL ACTIVITY 8

Projectile motion

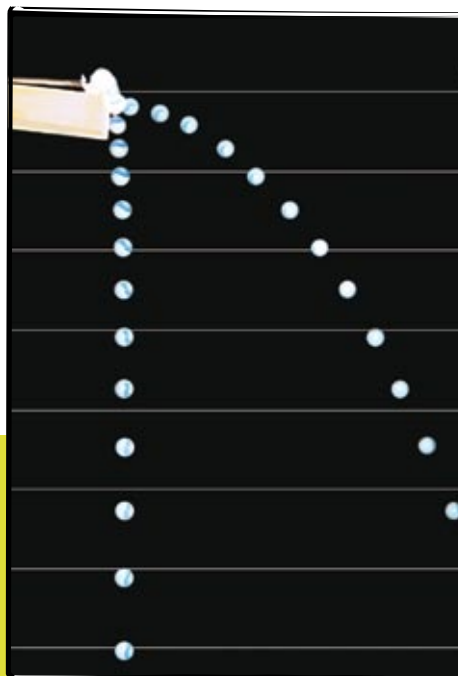


Figure 1.20 A multiflash photo of two golf balls released simultaneously. One ball was launched horizontally at 2.0 m s^{-1} while the other was released from rest. The projectile launched horizontally travels an equal horizontal distance during each flash interval, indicating that its horizontal velocity 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 9.8 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).

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.

Figure 1.21 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.

Physics file

It can be shown mathematically that the path of a projectile will be a parabola. Consider a projectile launched horizontally with speed v , and an acceleration down given by g . Let x and y be the horizontal and vertical displacements respectively.

At time t , the horizontal displacement is given by:

$$x = vt \quad (1)$$

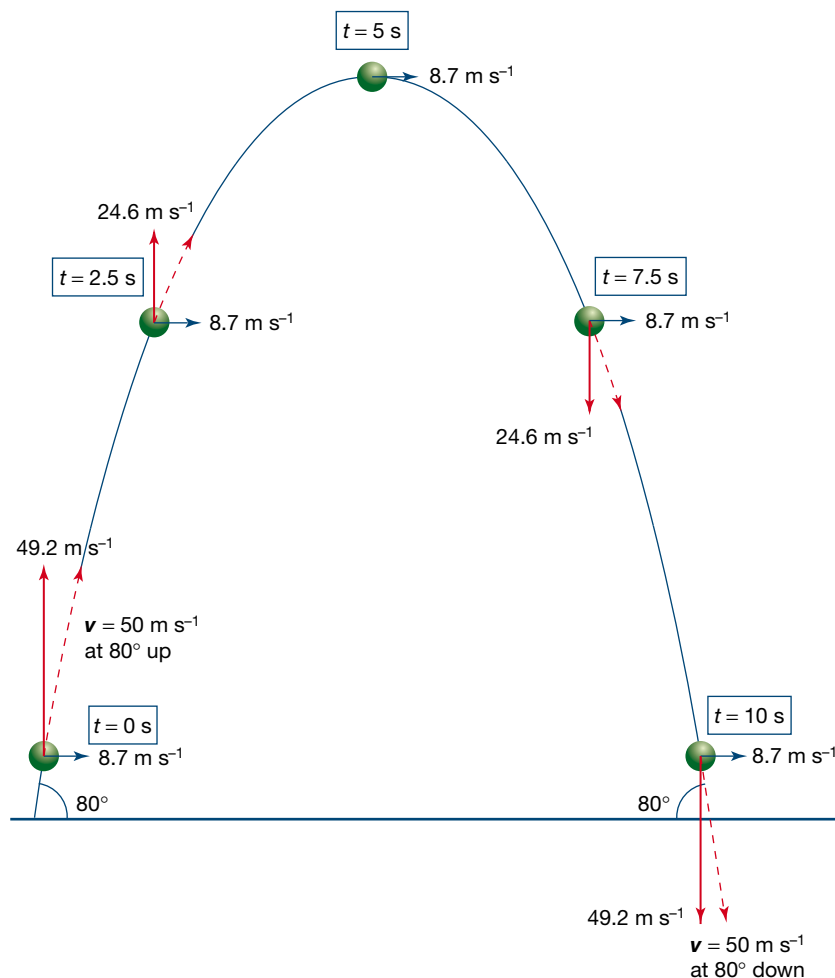
The vertical displacement at time t is:

$$y = \frac{1}{2}gt^2 \quad (2)$$

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

$$y = \frac{1}{2}g\left(\frac{x}{v}\right)^2 = \left(\frac{g}{2v^2}\right)x^2$$

Since g and v do not vary, $y \propto x^2$, which is the relationship for a parabola.



Tips for problems involving projectile motion

- Construct a diagram showing the motion and set the problem out clearly. Distinguish between information supplied for each component of the motion.
- In the *horizontal component*, the velocity, v , of the projectile is constant and so the only formula needed is $v = x/t$.
- For the *vertical component*, the projectile is moving with a constant acceleration (9.8 m s^{-2} down), and so the equations of motion for uniform acceleration must be used.
- In the vertical component, it is important to clearly specify whether up or down is the positive or negative direction, and use this consistently throughout the problem.

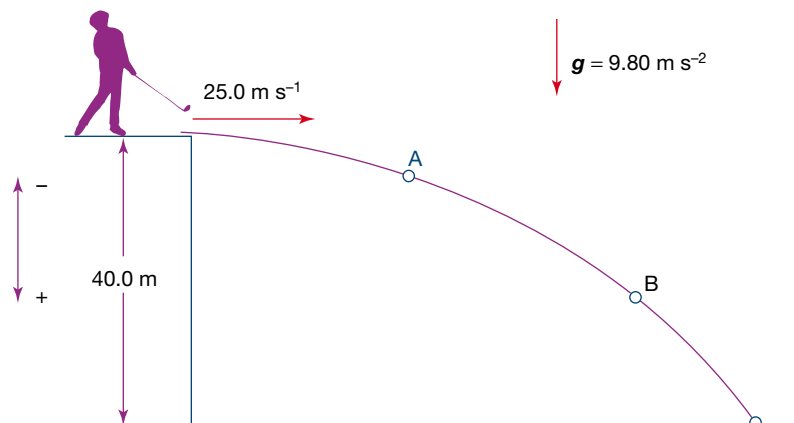
Worked example 1.5A

Horizontal launch

A golf ball of mass 150 g is hit horizontally from the top of a 40.0 m high cliff with a speed of 25.0 m s^{-1} . Assuming an acceleration due to gravity of 9.80 m s^{-2} and ignoring air resistance, calculate:

- the time that the ball takes to land
- the distance that the ball travels from the base of the cliff
- the velocity of the ball as it lands

- d the net force acting on the ball at points A and B
e the 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 positive direction:

$$u_v = 0, a = 9.80 \text{ m s}^{-2}, x = 40.0 \text{ m}, t = ?$$

Substituting in $x = ut + \frac{1}{2}at^2$ for the vertical direction only:

$$40.0 = 0 + 0.5 \times 9.80 \times t^2, \text{ so:}$$

$$t = \sqrt{40.0 / (0.5 \times 9.80)}$$

$$= 2.86 \text{ s}$$

The ball takes 2.86 s to reach the ground.

- b To find the horizontal distance travelled by the ball (i.e. the range of the ball), it is necessary to use the horizontal component.

$$u_h = 25.0 \text{ m s}^{-1}, t = 2.86 \text{ s}, x_h = ?$$

$$x_h = u_h t$$

$$x_h = 25.0 \times 2.86$$

$$= 71.5 \text{ m}$$

The ball lands 71.5 m from the base of the cliff (i.e. the range of the ball is 71.5 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:

$$u_v = 0, a = 9.80 \text{ m s}^{-2}, x = 40.0 \text{ m}, t = 2.86 \text{ s}$$

Substituting in $v = u + at$ for the vertical direction only:

$$v = 0 + (9.80 \times 2.86)$$

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

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:

$$v = \sqrt{25.0^2 + 28.0^2}$$

$$= \sqrt{1409}$$

$$= 37.5 \text{ m s}^{-1}$$

The angle at which it lands can be found by using trigonometry:

$$\tan \theta = \frac{28.0}{25.0}$$

$$= 1.12$$

$$\theta = 48.2^\circ$$

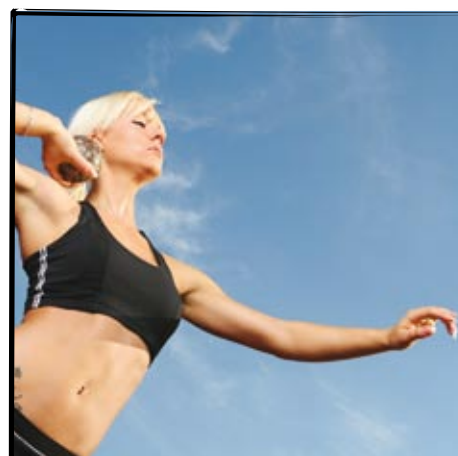
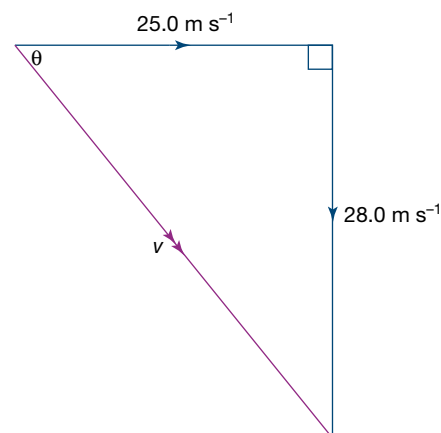


Figure 1.22 Taller athletes have an advantage in the shot-put. They launch the shot from a greater height and so will achieve a greater distance. Recent female world champions have been around 2.0 m tall.



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

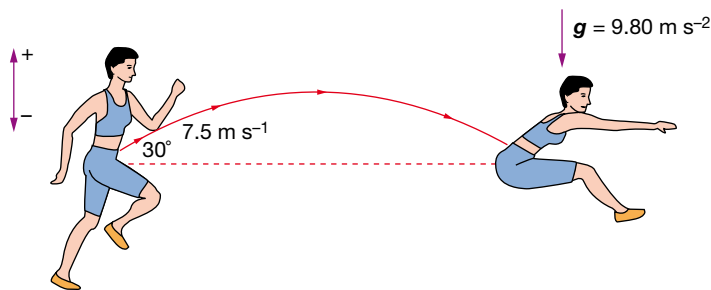
- d** If air resistance is ignored, the only force acting on the ball throughout its flight is its weight. Therefore the net force that is acting at point A and point B (and everywhere else!) is:

$$\begin{aligned}\Sigma F &= F_g = mg \\ &= 0.150 \times 9.80 \\ &= 1.47 \text{ N down}\end{aligned}$$

- 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 1.5B

Launch at an angle



A 65 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 g as 9.80 m s^{-2} , calculate:

- the horizontal component of the initial velocity
- the vertical component of the initial velocity
- the velocity when at the highest point
- the maximum height gained by the athlete
- the total time for which the athlete is in the air
- the horizontal distance travelled by the athlete's centre of mass (assuming that it returns to its original height)
- the 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 diagram, the horizontal component, u_h , of the athlete's initial velocity is:

$$\begin{aligned}u_h &= 7.50 \times \cos 30.0^\circ \\ &= 6.50 \text{ m s}^{-1} \text{ to the right}\end{aligned}$$

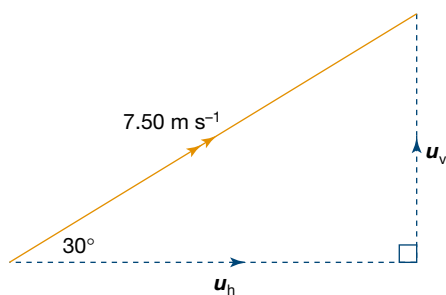
This remains constant throughout the jump.

- b** Again referring to the diagram, the vertical component, u_v , of the initial velocity of the athlete is:

$$\begin{aligned}u_v &= 7.50 \times \sin 30.0^\circ \\ &= 3.75 \text{ m s}^{-1} \text{ upwards}\end{aligned}$$

- 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 (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 (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:

$$u_v = 3.75 \text{ m s}^{-1}, v = 0, a = -9.80 \text{ m s}^{-2}, x = ?$$

$$v^2 = u^2 + 2ax$$

$$0 = 3.75^2 + [2 \times -9.80 \times x]$$

$$x = 0.717 \text{ m}$$

i.e. the centre of mass of the athlete rises by a maximum height of 72 cm.

- e** As the motion is symmetrical, 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:

$$u_v = 3.75 \text{ m s}^{-1}, v = 0, a = -9.80 \text{ m s}^{-2}, t = ?$$

$$v = u + at$$

$$0 = 3.75 + (-9.80 \times t)$$

$$t = 0.383 \text{ s}$$

The time for the complete flight is double the time to reach maximum height, i.e. total time in the air: $\Sigma t = 2 \times 0.383 = 0.766 \text{ s}$.

- 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 0.766 s and so:

$$t = 0.766 \text{ s}, v = 6.50 \text{ m s}^{-1}, x = ?$$

$$v = \frac{x}{t}, \text{ so}$$

$$x = v \times t$$

$$= 6.50 \times 0.766$$

$$= 4.98 \text{ m}$$

i.e. the athlete jumps a horizontal distance of 4.98 m.

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

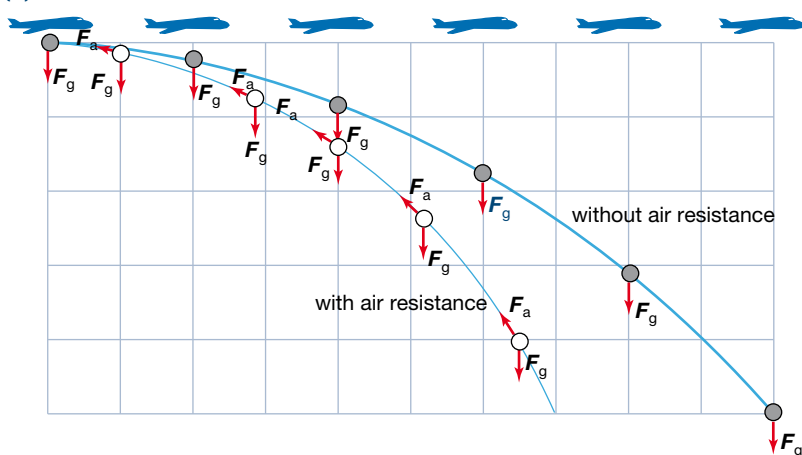
Physics file

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. Energy will be discussed in detail in Chapter 2.

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.

[a]



[b]

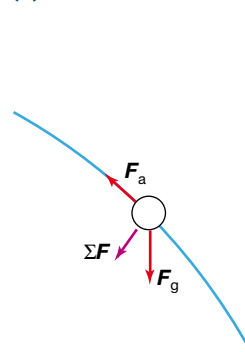


Figure 1.23 [a] The path of a food parcel dropped from a plane. If the plane maintains a constant speed and in the absence of air resistance, the parcel will fall in a parabolic path and remain directly below the plane. Air resistance makes the parcel fall more slowly, over a shorter path. [b] When air resistance is acting, the net force on the parcel is not vertically down.

Figure 1.23 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, F_g , and air resistance, F_a . Therefore, the resultant force, ΣF , 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.

Physics in action

Why ballet dancers seem to float in the air

The grande 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.

(a)



(b)

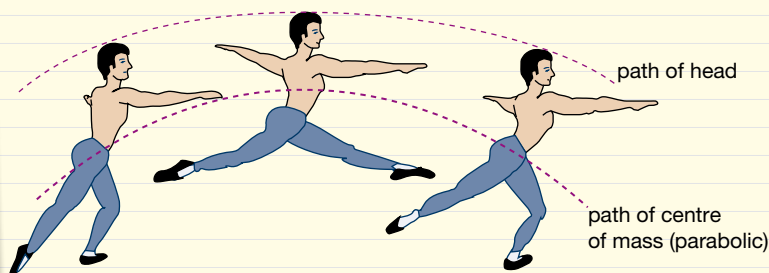


Figure 1.24 (a) Performing the grande jeté. (b) As the position of the centre of mass moves higher in the body, the head of the dancer follows a flatter path and this gives the audience the impression of a graceful floating movement.



1.4 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 its weight, i.e. the force of gravity, F_g or W . This results in the projectile having a vertical acceleration of 9.8 m s^{-2} down during its flight.
- The horizontal speed 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 as the vertical component 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.

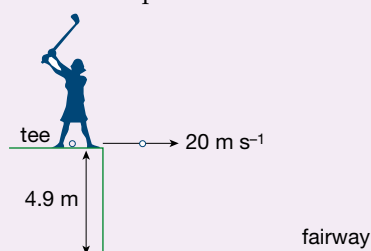


1.4 questions

Projectile motion

For the following questions, assume that the acceleration due to gravity is 9.8 m s^{-2} and ignore the effects of air resistance unless otherwise stated.

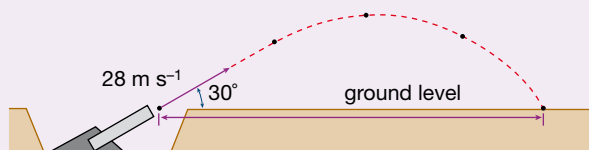
- 1 A golfer practising on a range with an elevated tee 4.9 m above the fairway is able to strike a ball so that it leaves the club with a horizontal velocity of 20 m s^{-1} .
 - a How long after the ball leaves the club will it land on the fairway?
 - b What horizontal distance will the ball travel before striking the fairway?
 - c What is the acceleration of the ball 0.50 s after being hit?
 - d Calculate the speed of the ball 0.80 s after it leaves the club.
 - e With what speed will the ball strike the ground?



- 2 A bowling ball of mass 7.5 kg travelling at 10 m s^{-1} rolls off a horizontal table 1.0 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 senior physics class conducting a research project on 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 m s^{-1} at an angle of 30° to the horizontal.



- 3 Calculate the horizontal component of the velocity of the ball:
 - a initially
 - b after 1.0 s
 - c after 2.0 s.

- 4 Calculate the vertical component of the velocity of the ball:
 - a initially
 - b after 1.0 s
 - c after 2.0 s.
- 5
 - a At what time will the ball reach its maximum height?
 - b What is the maximum height that is 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 minimum speed of the ball 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 g is thrown with an initial velocity of 16 m s^{-1} at an angle θ to the horizontal. When the ball reaches its maximum height, its kinetic energy is 16 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 1.0 s?
 - e What is the displacement of the ball after 1.0 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:
 - a the speed at which she leaves the ramp
 - b the maximum height above the end of the ramp that she reaches
 - c the time for which she is in mid-air.

chapter review

For the following questions, assume that the acceleration due to gravity is 9.8 m s^{-2} and ignore the effects of air resistance unless otherwise stated.

The following information applies to questions 1 and 2.

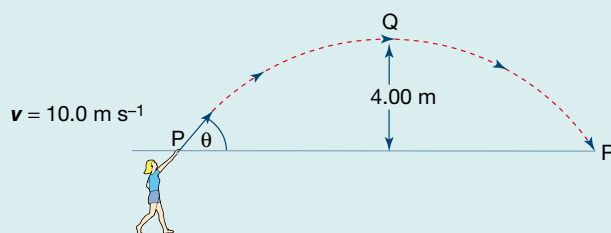
David is jumping on his trampoline. His mass is 25 kg and he lands vertically at 3.5 m s^{-1} before rebounding vertically at 3.0 m s^{-1} .

- 1
 - a What is David's change of speed as he bounces?
 - b What is David's change of velocity as he bounces?
- 2 Each time David bounces, the trampoline exerts a force on him.
 - 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 that David bounces, he exerts a downwards force on the trampoline. Which one of the following statements is correct?
 - A The force that David exerts on the trampoline is always greater than the force that the trampoline exerts on David.
 - B The force that David exerts on the trampoline is always less than the force that the trampoline exerts on David.
 - C The force that David exerts on the trampoline is at times greater than and at other times less than the force that the trampoline exerts on David.
 - D The force that David exerts on the trampoline is always equal to the force that the trampoline exerts on David.
- 3 An Olympic archery competitor tests a bow by firing an arrow of mass 25 g vertically into the air. The arrow leaves the bow with an initial vertical velocity of 100 m s^{-1} .
 - a At what time will the arrow reach its maximum height?
 - b What is the maximum vertical distance that this arrow reaches?
 - c What is the acceleration of the arrow when it reaches its maximum height?
- 4 A motorcyclist is travelling at 60 km h^{-1} when she has to brake to a sudden stop. She skids and stops in a distance of 15 m . The combined mass of the bike and rider is 120 kg .
 - a What was the magnitude of her average acceleration as she stopped?
 - b How fast was she travelling after skidding for 1.5 s ?
 - c Determine the magnitude of the average retarding force (in kN) acting on her motorbike as it stopped.
- 5 An aeroplane is headed due north at 500 km h^{-1} in still air. Then the wind starts to blow at a speed of 100 km h^{-1} towards the west.

- a What is the speed of the plane relative to the ground now?
 - A 600 km h^{-1}
 - B 400 km h^{-1}
 - C 510 km h^{-1}
 - D 490 km h^{-1}
 - b The pilot wishes to travel due north at 500 km h^{-1} . In which direction and with what air speed should the pilot fly the plane to achieve this goal?
 - A 500 km h^{-1} north
 - B 510 km h^{-1} north
 - C 490 km h^{-1} at 11° true
 - D 510 km h^{-1} at 11° true
- 6 Two identical tennis balls X and Y are hit horizontally from a point 2.0 m above the ground with different initial speeds: ball X has an initial speed of 5.0 m s^{-1} while ball Y has an initial speed of 7.5 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 7–10.

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.



- 7 What is the value of the angle θ that the initial velocity vector makes with the horizontal?
- 8 What is the speed of the Vortex at point Q?
- 9 What is the acceleration of the Vortex at point Q?
 - A zero
 - B 9.8 m s^{-2} forwards
 - C 4.9 m s^{-2} down
 - D 9.8 m s^{-2} down
- 10 How far away is the Vortex when it reaches point R?

The following information applies to questions 11–15.

In a shot-put event a 2.0 kg shot is launched from a height of 1.5 m, with an initial velocity of 8.0 m s^{-1} at an angle of 60° to the horizontal.

- 11 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.
- 12 What is the speed of the shot when it reaches its maximum height?
- 13 What is the minimum kinetic energy of the shot during its flight?
- 14 What is the acceleration of the shot at its maximum height?
- 15 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°
B 30°
C 45°
D 60°

The following information applies to questions 16–18.

During a volleyball match, Adam served the ball of mass 140 g from a height of 2.0 m. The ball was served at 12 m s^{-1} and reached a maximum height of 4.5 m above the court.



16 What was the initial kinetic energy of the volleyball?

- A 10 J
B 20 J
C 10 000 J
D 2.7 J

17 What was the initial gravitational potential energy of the volleyball?

18 How fast was the ball travelling at its highest point?

- A zero
B 9.6 m s^{-1}
C 9.8 m s^{-1}
D 92 m s^{-1}

The following information applies to questions 19 and 20.

A student at the Australian Institute of Sport was able to establish that during its flight, a 2.0 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 toss was launched at 7.5 m s^{-1} at an angle of 36° to the horizontal.

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

20 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?

Collisions and circular motion

This 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.



by the end of this chapter

you will have covered material from the study of collisions and circular motion, including:

- momentum and impulse
- conservation of momentum
- work and energy
- conservation of energy
- elastic and inelastic collisions
- circular motion in horizontal planes and vertical planes
- banked corners
- apparent weight.

2.1 Momentum and impulse

Momentum

When Newton outlined his second law of motion, he did not describe it in the way we use it today. He talked about a property of a moving object that he called its 'motion'. He said that when a force acts on an object for some time interval, its 'motion' would change. Today, we call this quantity the *momentum* of the object. In fact, Newton's first law could be stated as 'in the absence of unbalanced forces, the momentum of an object will be constant'.

Momentum is a property of moving bodies. A tennis ball has momentum as it flies through the air; skiers have momentum as they race down a slope; even a snail has momentum as it slowly slithers across a footpath. Stationary objects, no matter how massive, have no momentum.

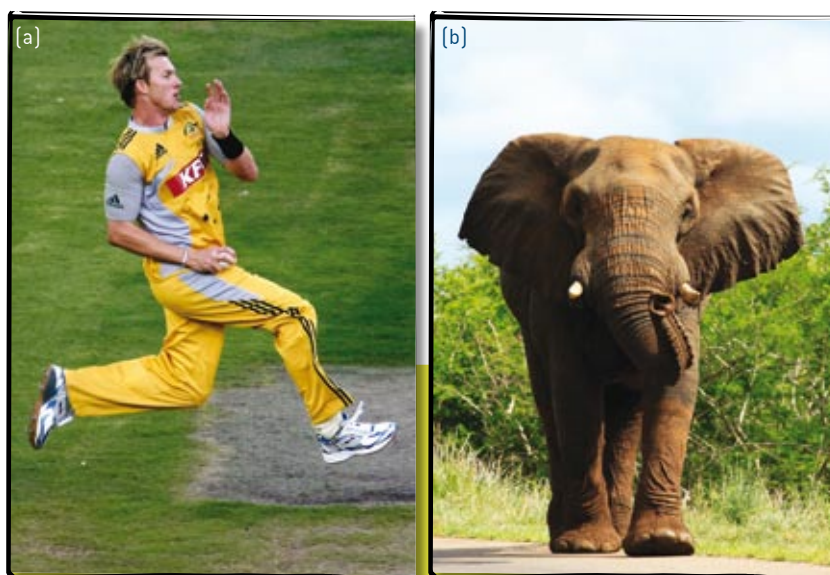


Figure 2.1 Which of these moving objects has more momentum: (a) the cricket ball when bowled by Brett Lee or (b) the elephant? The cricket ball moves a lot faster, but has much less mass. According to Newton, the object with less momentum is the one that is easier to stop. To determine the answer to the question, calculations are required.



MOMENTUM (p) is defined as the product of the mass and velocity of an object. Momentum is a vector quantity, and its direction is that of the velocity. The units of momentum are kilogram metres per second (kg m s^{-1}):

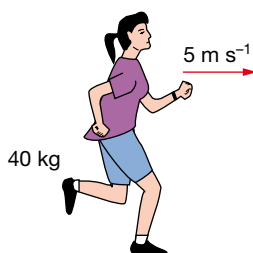
$$p = \text{mass} \times \text{velocity or } p = mv$$

where p = momentum (kg m s^{-1})

m = mass (kg)

v = velocity (m s^{-1})

(a) W ← → E



(b) W ← → E

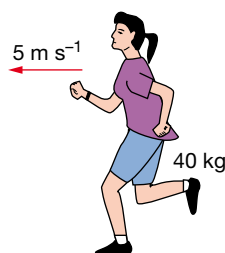


Figure 2.2 The speed of this girl is the same in both diagrams, but her momentum is different.

(a) The momentum of the girl is 200 kg m s^{-1} east.
(b) Her momentum is now 200 kg m s^{-1} west.

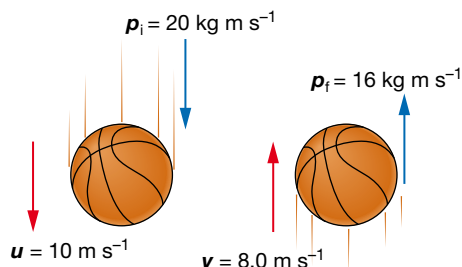


Figure 2.3 The momentum of this basketball has changed from 20 kg m s^{-1} down to 16 kg m s^{-1} up. The change in momentum can be determined by performing a vector subtraction: $\Delta p = p_f - p_i$.

Physics file

Since impulse can be expressed in terms of a momentum change, the units for momentum (kg m s^{-1}) and impulse (N s) must be equivalent. This can be shown by using Newton's second law.

Given that $1 \text{ N} = 1 \text{ kg m s}^{-2}$

(from $\Sigma F = ma$)

it follows that $1 \text{ N s} = 1 \text{ kg m s}^{-2} \times \text{s}$

i.e. $1 \text{ N s} = 1 \text{ kg m s}^{-1}$

Even though the units are equivalent, it is advisable that they be used with the appropriate quantities as a reminder of the quantity that is being dealt with.

The newton second (N s) is the product of a force and a time interval and so should be used with impulse.

The kilogram metre per second (kg m s^{-1}) is the product of a mass and a velocity and so should be used with momentum quantities.

Even so, it is not uncommon to see newton seconds used as the units of momentum, nor is this incorrect.

It is important to remember that the momentum of an object depends on the *direction* in which it is travelling. For example, a girl of mass 40 kg running towards the east at 5 m s^{-1} has a different momentum from the same girl running towards the west at 5 m s^{-1} .

Change in momentum

The momentum of an object will change when it experiences an unbalanced force, for example, as the result of a collision or interaction with another object. It is often necessary to find the change in momentum of the object.



CHANGE IN MOMENTUM (Δp) is defined as the difference between the final momentum and the initial momentum.

Change in momentum = final momentum – initial momentum

$$\Delta p = p_f - p_i$$

Since momentum is a vector quantity, a *vector subtraction* must be performed to determine a change in momentum.

Consider the example of a basketball of mass 2.0 kg that falls vertically onto a floor with a speed 10 m s^{-1} and rebounds at 8.0 m s^{-1} . Taking up as being the positive direction, the change in momentum of the basketball is:

$$\begin{aligned}\Delta p &= p_f - p_i \\ &= (+16) - (-20) \\ &= 36 \text{ kg m s}^{-1}\end{aligned}$$

i.e. the change in momentum of the basketball is 36 kg m s^{-1} up.

Impulse

As Newton worked out over 300 years ago, momentum changes are caused by unbalanced forces. Consider the example of a mass, m , that is acted upon by a net force, ΣF , for a time interval, Δt . The mass will accelerate while this unbalanced force is acting, as described by Newton's second law:

$$\begin{aligned}\Sigma F &= ma \\ \therefore \Sigma F &= \frac{m\Delta v}{\Delta t} \\ \therefore \Sigma F\Delta t &= m\Delta v\end{aligned}$$

The term $m\Delta v$ is the momentum change, Δp , of the mass while the net force is acting. The product of the net force, ΣF , and the time interval, Δt , that it acts for is called the net *impulse* of the net force, i.e. net impulse = $\Sigma F\Delta t$.

Thus it can be seen that the net impulse an object experiences is equal to its change in momentum, i.e. net impulse = change in momentum.



Net IMPULSE is given by $\Sigma F\Delta t = m\Delta v = \Delta p$ = change in momentum where ΣF = net force (N)

Δt = time interval (s)

m = mass (kg)

Δv = change in velocity (m s^{-1})

Force–time graphs

When a constant force is acting on an object, the impulse equation can be used to work out the change in momentum. In most real-life situations, however, the force will vary in size during the interaction. When this happens, force–time (F – t) graphs can be used to determine the impulse of a force.



Impulse is given by the area under a force–time graph.

We will now use force–time graphs to investigate an interesting aspect of motion—why stopping suddenly is more painful than stopping gradually!

Consider the example of Siobhan, of mass 50 kg, who is skiing horizontally across snow at 6.0 m s^{-1} . Siobhan, therefore, has an initial momentum of 300 kg m s^{-1} . In order to improve her understanding of forces and impulse, Siobhan decides to experiment by stopping in two different ways. First, she falls to the snow and gradually slides to rest in 4.0 s. Then, in an act of unexplained bravado, she crashes into a tree and stops abruptly in just 0.10 s. The stopping forces acting on Siobhan in both cases are shown in Figure 2.5.

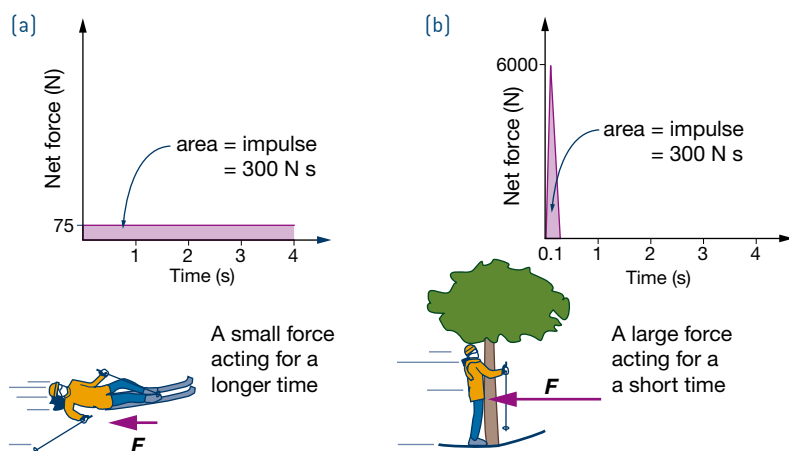


Figure 2.5 These graphs show the force that acts to stop the skier when (a) she slides to rest in the snow and (b) she crashes into a tree. The impulse (and momentum change) that the skier has experienced is the same in both cases.

In both cases, Siobhan comes to rest and so loses 300 kg m s^{-1} of momentum. In other words, the impulse of the force is 300 N s in both cases. This is also the result of finding the *area* under each of the \mathbf{F} – t graphs.

However, an important difference is the way in which she has stopped. When sliding through the snow, Siobhan comes to rest in 4.0 s and so a relatively small average force of 75 N is needed to cause her to lose all of her initial momentum. But the tree makes Siobhan come to rest in just 0.10 s. As seen in Figure 2.5, the peak force must necessarily be relatively large, making this a rather painful and possibly dangerous activity!

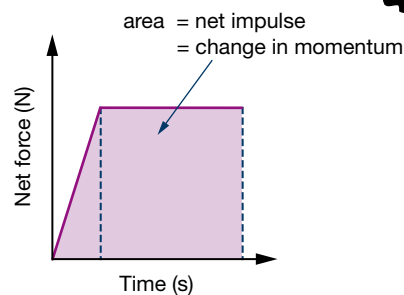


Figure 2.4 The area under a force–time graph gives the impulse produced by the force or the momentum change experienced by the object.

Physics file

It is important to note that impulse and force are not the same thing. A force is the push or pull that acts on a body. Impulse is the product of the force that acts and the time that it acts for and gives a measure of the resulting change in momentum. A small force acting over a long time interval can produce the same momentum change as a large force acting for a short time.

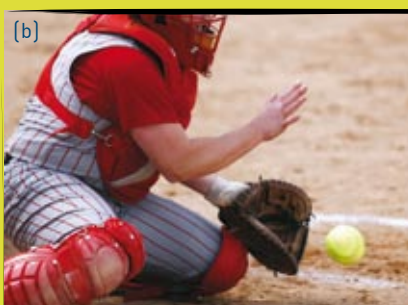


Figure 2.6 (a) The landing mat extends the time over which the athlete comes to rest, reducing the size of the stopping forces. If he missed the mat and landed on the ground, the forces would be larger, but his momentum change would be the same. (b) Helmets and vests contain padding that extends the time over which the softball comes to rest in a collision with the catcher, thereby reducing the size of the stopping force. (c) Wicket-keepers allow their hands to 'give' when keeping to a bowler. This extends the stopping time and reduces the stopping force.



Figure 2.8 It's not how fast you are travelling that matters—it's how fast you stop. In many sports and activities, care is taken to ensure that the participants or objects stop gradually rather than suddenly. This rider in the Australian MotoGP at Phillip Island was travelling at over 200 km h^{-1} as he came off. The gravel or 'kitty litter' alongside the track caused him to lose his momentum gradually by applying a relatively small retarding force on his body as he tumbled and slid to a stop. He walked away unharmed.

Worked example 2.1A

A ball of 'slime' of mass 100 g is thrown horizontally at a wall. It is travelling at 15 m s^{-1} south as it hits the wall and rebounds at 3.0 m s^{-1} north. The contact with the wall lasts for 20 ms . Taking north as the positive direction, calculate:

- the change in momentum of the slime
- the net impulse that acts on the slime as it rebounds
- the average force that the wall exerts on the slime
- the average force that the slime exerts on the wall.

Solution

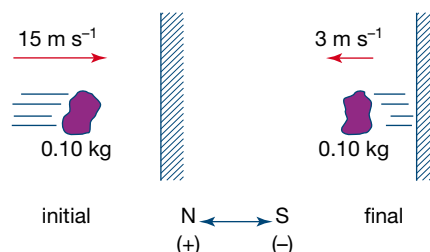


Figure 2.7 The net impulse from the wall causes the slime ball to experience a change in momentum. The slime has 1.5 kg m s^{-1} of momentum as it reaches the wall and rebounds with 0.3 kg m s^{-1} of momentum. Its change in momentum can be found by vector subtraction.

- Change in momentum = final momentum – initial momentum

$$\Delta p = p_f - p_i$$

$$= +0.30 - (-1.5)$$

$$= +1.8 \text{ kg m s}^{-1}$$

i.e. the momentum change of the slime ball as it bounces off the wall is 1.8 kg m s^{-1} north.
- The net impulse that the slime experiences is equal to the change in momentum of the slime. Therefore the impulse is 1.8 N s north.
- The force that the wall exerts on the slime will actually vary over the contact time, but it is possible to determine the size of the average force that acts during this interaction. The impulse that the slime experiences is equal to its change in momentum.

$$\Sigma F \Delta t = \Delta p$$

$$\Sigma F \times 0.020 = 1.8 \text{ kg m s}^{-1} \text{ north}$$

$$\Sigma F = 90 \text{ N to the north}$$

i.e. the wall exerts an average net force of 90 N north on the slime.
- According to Newton's third law, the force that the slime exerts on the wall is equal but opposite to the force that the wall exerts on the slime, i.e. the slime exerts an average net force of 90 N south on the wall.

Physics in action

Car safety and crumple zones

Worldwide, automobile accidents are responsible for over 2 million deaths each decade. Many times this number of people are injured. One way of reducing the toll is to design safer vehicles. Modern cars employ a variety of safety features that help to improve the occupants' survival chances in an accident. Some of these safety features are the antilock braking system (ABS), electronic stability control (ESC), inertia reel seatbelts, variable-ratio response steering systems, collapsible steering columns, head rests, shatterproof windscreen glass, padded dashboards, front and side air bags, front and rear crumple zones and a rigid passenger compartment.

Some prototype cars today are equipped with collision avoidance systems. These have laser or infrared sensors that advise the driver of hazardous situations, and even take over the driving of the car in order to avoid accidents! Consider the example of the driver of a car that crashes into a tree at 60 km h^{-1} (16.7 m s^{-1}). If the driver has a mass of 90 kg , then the momentum of the driver is:

$$p = m \times v = 90 \times 16.7 = 1500 \text{ kg m s}^{-1}$$

As a result of this collision, the driver will lose all of this momentum as the car comes suddenly to a stop. However, the



Figure 2.9 Cars are designed with weak points in their chassis at the front and rear that enable them to crumple in the event of a collision. This extends the time over which the cars come to rest and so reduces the size of the forces acting on the occupants.

impulse experienced by the driver is the same whether the stop is sudden or gradual. In either case, the impulse is -1500 N s . So the idea of safety features such as inertia reel seatbelts, collapsible steering columns, padded dashboards, air bags and crumple zones is not to reduce the size of the impulse, but to reduce the size of the forces that act to bring the driver to a stop. Automotive engineers strive to achieve this by extending the time over which the driver loses momentum.

Crumple zones

A popular misconception among motorists is that cars would be much safer if they were sturdier and more rigid. Drivers often complain that cars seem to collapse too easily during collisions, and that it would be better if cars were structurally stronger—more like an army tank. In fact, cars are specifically designed to crumple to some extent. This makes them safer and actually reduces the seriousness of injuries suffered in car accidents.

Army tanks are designed to be extremely sturdy and rigid vehicles. They are able to withstand the effect of



Figure 2.10 The Australian New Car Assessment Program (ANCAP) assesses the crashworthiness of new cars. This car has just crashed at 50 km h^{-1} into a 5-tonne concrete block. The crumpling effect can clearly be seen.

collisions without suffering serious structural damage. If a tank travelling at 50 km h^{-1} crashed into a solid obstacle, the tank would be relatively undamaged. However, its occupants would very likely be killed. This is because the tank has no 'give' in its structure and so the tank and its occupants would stop in an extremely short time interval. The occupants would lose all of their momentum in an instant, which means that the forces acting on them must necessarily be extremely large. These large forces would cause the occupants of the tank to sustain very serious injuries, even if they were wearing seatbelts.

Motor cars today have strong and rigid passenger compartments; however, they are also designed with non-rigid sections such as bonnets and boots that crumple if the cars are struck from the front or rear. The chassis contains members that have grooves or beads cast into them. In a collision, these beads act as weak points in the members, causing them to crumple in a concertina shape.

This 'concertina' effect allows the front or rear of the car to crumple, extending the time interval over which the car and its occupants come to a stop. This stopping time is typically longer than 0.1 s in a 50 km h^{-1} crash. Because the occupants' momentum is lost more gradually, the peak forces that act on them are smaller and so the chances of injury are reduced.



2.1 summary

Momentum and impulse

- The momentum, p , of an object is the product of its mass and velocity. Momentum is a vector quantity:

$$p = mv$$

The units of momentum are kilogram metres per second (kg m s^{-1}).

- When a net force acts on an object for a given time, the object will experience a change in momentum, Δp .
- The product of the net force, ΣF , acting on an object and the time interval, Δt , that this force acts for is called

the net impulse of the net force. Impulse is a vector quantity:

$$\text{Impulse} = \Sigma F \Delta t = m \Delta v = \Delta p$$

The units of impulse are newton seconds (N s).

- Impulse can be determined by finding the area under a force–time graph.
- The size of the net force that is acting on an object is directly related to the rate at which the momentum of the object changes.



2.1 questions

Momentum and impulse

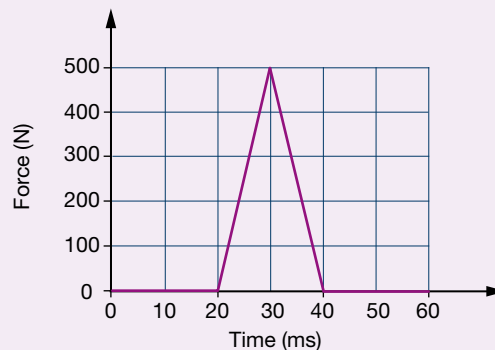
- Which has more momentum—a mosquito buzzing along or a stationary bus?
- Pavithra and Michelle are having a race. Pavithra, whose mass is 45 kg, is running at 3.5 m s^{-1} . Michelle has a mass of 60 kg and runs at a speed of 2.5 m s^{-1} . Which girl has a greater amount of momentum? Use calculations to determine the answer.
- Calculate the momentum of these objects:
 - an electron of mass $9.1 \times 10^{-31} \text{ kg}$ moving at $3.0 \times 10^7 \text{ m s}^{-1}$
 - the Earth, of mass $6.0 \times 10^{24} \text{ kg}$ and orbital speed 30 km s^{-1}
 - a mouse of mass 100 g running at 2.3 m s^{-1} .
- During training, a swimmer turns at the end of a lap and swims back with equal speed. Which one of these statements is correct?
 - The swimmer has more momentum after the turn.
 - The swimmer has less momentum after the turn.
 - The swimmer has the same momentum after the turn.
 - The swimmer has different momentum after the turn.
- Vijay is standing on a high fence and about to jump down to the ground.
 - When Vijay lands, he automatically bends his knees as he reaches the ground. Why does he do this? Explain in terms of physics principles that you have learnt.
 - A friend tries to talk Vijay into keeping his legs stiff as he lands. What would happen if Vijay took this advice?

The following information applies to questions 6 and 7.

A rubber ball of mass 80 g bounces vertically on a concrete floor. The ball strikes the floor at 10 m s^{-1} and rebounds at 8.0 m s^{-1} .

- What is the change in speed of the ball as it bounces?
 - What is the change in velocity of the ball as it bounces?
 - What is the change in momentum of this ball during its bounce?
 - Calculate the impulse acting on the ball during its bounce.
- The time of contact between the ball and the floor during the bounce was 0.050 s.
 - Calculate the average net force acting on the ball during its contact with the floor.
 - Calculate the average force that the floor exerts on the ball.
 - Calculate the average force that the ball exerts on the floor.

- During a tennis match, a tennis ball of mass 100 g is served. Assume that the initial speed of the ball is zero. The racquet exerts a force on the ball as shown in the diagram below.



- What is the maximum force that the racquet exerts on the ball?
 - What is the maximum force that the ball exerts on the racquet?
 - Calculate the net impulse delivered to the ball.
 - What is the change in momentum of the ball?
 - Calculate the speed with which this ball was served.
- Explain why protective padding placed on netball posts reduces the likelihood of injuries for players who collide with them.
 - In an experiment to test the benefits of incorporating a crumple zone into the design of a car, two otherwise identical cars are modified so that, on collision, the passenger compartment of each car will take a different time to stop. The cars, A and B, containing identical dummies of mass 100 kg, are driven into a solid concrete wall at 20 m s^{-1} . The results are as follows. The passenger compartment of car A, with the larger crumple zone, took 0.40 s to stop after the collision. Car B took only 0.10 s to stop after the collision. The crash dummies were firmly attached to the passenger compartments of their respective vehicles by extremely tight-fitting seatbelts.
 - In which car did the dummy experience the greater:
 - momentum change?
 - greater net impulse?
 - stopping force?
 - Calculate the average force exerted by the seatbelts on each crash dummy during the collision.
 - What conclusion can you make concerning the benefit of crumple zones?
 - Explain why seatbelts are designed to stretch and allow a passenger to move a short distance during a collision rather than providing unyielding restraint.

2.2 Conservation of momentum

Collisions are an inescapable aspect of everyday life and the workings of the Universe. In this section, the term ‘collision’ describes interaction situations in which objects come together and exert action/reaction forces on each other. A collision does not necessarily require actual contact between the objects—just an interaction based on force, as in two protons approaching each other and being repelled.

Consider a collision between a bowling ball and one of the pins (Figure 2.11). The ball has a mass of m_b and an initial velocity of \mathbf{u}_b . It collides with pin of mass m_p that has an initial velocity \mathbf{u}_p of zero. After the collision, the ball has a reduced velocity of \mathbf{v}_b and the pin is moving with a higher velocity of \mathbf{v}_p . According to Newton’s third law, equal but opposite forces act on the ball and pin during the actual contact between them.

The ball experiences a force, \mathbf{F}_b , due to its collision with the pin. This causes the ball to slow down slightly. Similarly, the pin experiences an equal but opposite force, \mathbf{F}_p , due to its collision with the ball; that is, $\mathbf{F}_b = -\mathbf{F}_p$.

Now, from Newton’s second law it follows that $m_b \times \mathbf{a}_b = m_p \times \mathbf{a}_p$, so:

$$m_b \left(\frac{\mathbf{v}_b - \mathbf{u}_b}{\Delta t_b} \right) = -m_p \left(\frac{\mathbf{v}_p - \mathbf{u}_p}{\Delta t_p} \right)$$

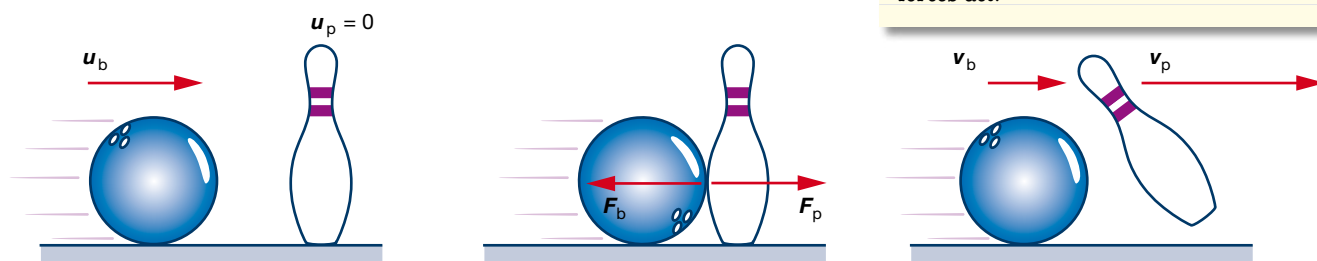


Figure 2.11 When a bowling ball collides with a tenpin, they exert equal but opposite forces on each other. These forces cause the ball to lose some momentum and the pin to gain an equal amount of momentum.

The time interval for which the ball and pin are in contact with each other is equal, i.e. $\Delta t_b = \Delta t_p$, so:

$$\begin{aligned} m_b(\mathbf{v}_b - \mathbf{u}_b) &= -m_p(\mathbf{v}_p - \mathbf{u}_p) \\ m_b \mathbf{v}_b - m_b \mathbf{u}_b &= -m_p \mathbf{v}_p + m_p \mathbf{u}_p \end{aligned}$$

and thus

$$m_b \mathbf{u}_b + m_p \mathbf{u}_p = m_b \mathbf{v}_b + m_p \mathbf{v}_p$$

The terms on the left of the equation give the momentum of the ball and the pin *before* the collision. The terms on the right give the total momentum of the objects *after* the collision. The expression says that *the total momentum before the collision is equal to the total momentum after the collision*. This is known as the *law of conservation of momentum*.



The **LAW OF CONSERVATION OF MOMENTUM** states that, in any collision or interaction between two or more objects in an isolated system, the total momentum of the system will remain constant; that is, the total initial momentum will equal the total final momentum:

$$\Sigma \mathbf{p}_i = \Sigma \mathbf{p}_f$$

This is a very powerful law and applies to any type of interaction in the Universe involving any number of objects. It is important to understand

Physics file

When Newton analysed collisions, he assumed that when a force acts on an object, the ‘motion’ of the object would change. He said ‘The quantity of motion...suffers no change from the actions of bodies among themselves’. The ‘motion’ is what we now call momentum. During any collision, the time intervals that the forces act for are equal. The forces that the objects exert on each other are equal in magnitude but opposite in direction, so the momentum change of each object will also be equal but opposite. Newton’s third law could be expressed as: ‘If the only forces acting are action/reaction forces, the momentum gained by one object will be equal to the momentum lost by the other’. In other words, during a collision, momentum is conserved if no external forces act.



PRACTICAL ACTIVITY 9

Conservation of momentum in collisions

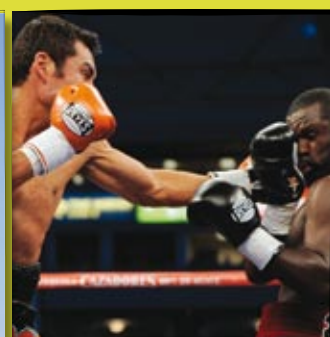


Figure 2.12 Only one of these interactions can be considered to be an isolated system. Can you work out which one? What are the external forces acting in the others?

that conservation of momentum applies to an *isolated* system. An isolated system is one where the collision involves only *internal* forces (i.e. the action/reaction forces on the objects involved in the collision). If the tennis ball crashed into the barrier and was stopped, this stopping force would be considered to be an *external* force; it is not a force that was considered within the system being analysed (the collision between the ball and pin). On Earth, perfectly isolated systems cannot exist because of the presence of gravitational and frictional forces. Only one of the collisions shown in Figure 2.12 could be considered to have occurred in an almost isolated system. See if you can identify it.

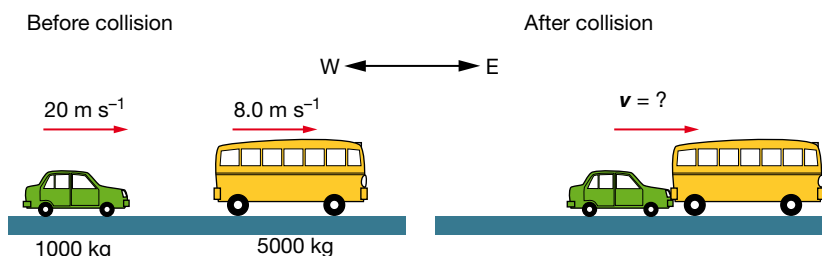
Physics file

Conservation of momentum helped scientists discover the neutrino. In the 1920s it was observed that in a beta decay, a nucleus emitted a beta particle (an electron emitted from a radioactive nucleus). However, when the nucleus recoiled, it was not in the opposite direction to the emitted electron. Thus, the momentum of these particles did not comply with the law of conservation of momentum. In 1930 Wolfgang Pauli determined that another particle must also have been emitted. This particle, the neutrino, was not detected experimentally until 1956. As you read this, billions of neutrinos are passing through your body and the Earth!

Worked example 2.2A

In a rear-end collision on a freeway, a car of mass 1.0×10^3 kg travelling east at 20 m s^{-1} crashes into the back of a bus of mass 5.0×10^3 kg heading east at 8.0 m s^{-1} . If the car and the bus lock together on impact, and friction is ignored, calculate:

- the final common velocity of the vehicles
- the change in momentum experienced by the car
- the change in momentum that the bus has experienced.



Solution

- In this problem, east will be treated as positive. Using conservation of momentum:

$$\Sigma p_i = \Sigma p_f$$

$$m_c u_c + m_b u_b = (m_c + m_b) v$$

$$(1000 \times +20) + (5000 \times +8.0) = 6000 \times v$$

$$60\,000 = 6000 \times v$$

$$v = +10 \text{ m s}^{-1}, \text{ i.e. } 10 \text{ m s}^{-1} \text{ east}$$

- The change in momentum of the car is given by:

$$\Delta p_c = p_c(f) - p_c(i)$$

$$= 10\,000 - 20\,000 = -10\,000 \text{ kg m s}^{-1}, \text{ i.e. } 1.0 \times 10^4 \text{ kg m s}^{-1} \text{ west}$$

- The change in momentum of the bus is equal to:

$$\Delta p_b = p_b(f) - p_b(i)$$

$$= 50\,000 - 40\,000 = +10\,000 \text{ kg m s}^{-1}, \text{ i.e. } 1.0 \times 10^4 \text{ kg m s}^{-1} \text{ east}$$

Note that the car has lost $10\,000\text{ kg m s}^{-1}$ and the bus has gained $10\,000\text{ kg m s}^{-1}$ of momentum during the collision. This is not a coincidence. If momentum is to be conserved in a collision or interaction between two objects, what one object loses in momentum, the other object must gain.

Worked example 2.2B

A 175 kg wrestler running east with a speed of 3.0 m s^{-1} crashes into an opponent of mass 100 kg running in the opposite direction at 5.0 m s^{-1} . The two wrestlers collide while in mid-air and remain locked together after their collision.

- Calculate the final velocity of the wrestlers.
- Some time later, the 100 kg wrestler is hurled into the turnbuckle at 5.0 m s^{-1} and comes to a stop. Where has the momentum of this wrestler gone?

Solution

In this problem east will be treated as positive and west will be treated as negative.

$$\begin{aligned} \mathbf{a} \quad \Sigma p_i &= \Sigma p_f \\ p_i[175\text{ kg wrestler}] + p_i[100\text{ kg wrestler}] &= p_f[\text{wrestlers locked together}] \\ (175 \times 3.0) + (100 \times -5.0) &= (175 + 100)v_f \\ 525 - 500 &= 275v_f \\ v_f &= \frac{25}{275} \\ &= 0.091\text{ m s}^{-1} \end{aligned}$$

The final common velocity of the wrestlers is 0.091 m s^{-1} east.

- This momentum (500 kg m s^{-1}) has been transferred to the ring and the Earth, causing the ring and Earth to move slightly.

Physics file

Circus strongmen often perform a feat where they place a large rock on their chest, then allow another person to smash the rock with a sledgehammer. This might seem at first to be an act of extreme strength and daring. However, a quick analysis using the principle of conservation of momentum will show otherwise. Let us assume that the rock has a mass of 15 kg and that the sledgehammer of mass 3 kg strikes it at 5 m s^{-1} . Using conservation of momentum, we can show that the rock and sledgehammer will move together at just 1 m s^{-1} after the impact.

The large mass of the rock has dictated that the final common speed is too low to hurt the strongman. A more daring feat would be to use the sledgehammer to smash a pebble!



Physics in action

Rockets

Early in the 20th century an American scientist, Robert Goddard, suggested that a rocket could be sent to the Moon. This view was commonly opposed. People thought that the vacuum in space would give the rocket gases nothing to push against. Goddard set up an experiment in which a gun fired a blank cartridge in a vacuum chamber. Hot gases and some wadding were projected from the gun during the blast. The gun recoiled even though these expelled gases had nothing to push against. Rocket propulsion had been demonstrated.

Figure 2.13 A space shuttle is attached to giant fuel tanks at launch. As the fuel in these tanks is used, the tanks are discarded. The momentum of the expelled fuel gives the rocket the momentum it needs to accelerate from the ground. The space shuttle program will finish in 2010 and be replaced by the Orion program.



When a gun fires a bullet, a Newton's third law action/reaction pair of forces acts to propel the bullet and to make the gun recoil. The magnitude of the momentum of the gun after the explosion equals that of the bullet. Rockets work on the same principle, except that hot gas replaces the bullet. It seems odd that some hot gas can make a massive rocket move so fast, but you need to keep in mind that the gases are expelled at extremely high speed and that about 90% of the initial mass of the rocket is in fact fuel.

The space shuttle is basically a glider attached to several enormous fuel tanks!

The force, or thrust, acting on a rocket can be determined if the exhaust velocity and rate at which the fuel is consumed are known. From the impulse, $F\Delta t = m\Delta v$, so (ignoring the vector nature of the quantities):

$$F = \left(\frac{m}{\Delta t}\right) \times \Delta v$$



PRACTICAL ACTIVITY 10

Conservation of momentum in explosions

where $m/\Delta t$ is the rate at which fuel is used and Δv is the exhaust velocity relative to the rocket.

For example, the first stage of a Saturn V rocket used 15 000 kg of fuel each second, and expelled it at 2500 m s^{-1} relative to the rocket. The thrust produced by this rocket was:

$$\begin{aligned} F &= \left(\frac{m}{\Delta t}\right) \times \Delta v \\ &= 1.5 \times 10^4 \times 2.5 \times 10^3 \\ &= 3.8 \times 10^7 \text{ N} \end{aligned}$$

As you can see, the forces that the rocket vehicle and the exhaust fuel exert on each other are enormous. This thrust force is opposed by a much smaller force of gravity. If the fuel is expelled at a uniform rate, the thrust acting on the rocket will remain constant. However, this force acts on a decreasing mass as the fuel is consumed. Therefore (as described by Newton's second law), the acceleration of the rocket will increase as it uses its fuel. During a shuttle launch, the two solid rocket booster tanks are jettisoned after 2 min. These deploy parachutes to descend and are recovered about 250 km away. The large external tank is jettisoned 8 min into the mission and falls into the ocean.



2.2 summary

Conservation of momentum

- An isolated system is one in which only action/reaction forces are considered to be acting. External forces (e.g. friction and gravity) are either non-existent or ignored.
- In any collision in an isolated system, the law of conservation of momentum says that the total

momentum of the system is conserved. The total momentum before the collision is equal to the total momentum after the collision:

$$\Sigma p_i = \Sigma p_f$$

- In a two-body collision, the momentum lost by one body must be gained by the other.



2.2 questions

Conservation of momentum

The following information applies to questions 1 and 2. A sports car of mass $1.0 \times 10^3 \text{ kg}$ travelling east at 36 km h^{-1} approaches a station wagon of mass $2.0 \times 10^3 \text{ kg}$ moving west at 18 km h^{-1} .

- 1 a Calculate the momentum of the sports car.
b Calculate the momentum of the station wagon.
c Determine the total momentum of these vehicles.
- 2 These two vehicles now collide head-on on an icy stretch of road where there is negligible friction. The vehicles remain locked together after the collision.
a Calculate their common velocity after the collision.
b Where has the initial momentum of the vehicles gone?

- c Determine the change in momentum of the sports car.

- d Determine the change in momentum of the station wagon.

- 3 A 200 g snooker ball travelling with initial velocity 9.0 m s^{-1} to the right collides with a stationary ball of mass 100 g. If the final velocity of the 200 g ball is 3.0 m s^{-1} to the right, calculate the velocity of the 100 g ball after the collision.

- 4 A 1000 kg cannon mounted on wheels fires a 10.0 kg shell with a horizontal speed of 500 m s^{-1} . Assuming that friction is negligible, calculate the recoil velocity of the cannon.



5 An arrow of mass 100 g is fired with an initial horizontal velocity of 40 m s^{-1} to the right at an apple of mass 80 g that is initially at rest on a horizontal surface. When the arrow strikes the apple, the two objects stick together. What is the common velocity of the arrow and apple after the impact?

6 A railway water tanker is rolling freely along train tracks at 5.0 m s^{-1} . A worker opens a plug hole so that water gushes out the bottom of the tanker. Joe and Mary, who were watching, disagreed as to what would happen next. Joe said that since the tanker was losing mass, its speed would increase in order to comply with the law of conservation of momentum. Mary said that the speed of the cart would not change and momentum would be conserved. Who was correct? Explain.

The following information applies to questions 7–9.

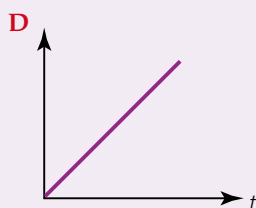
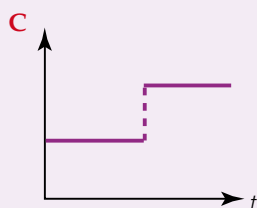
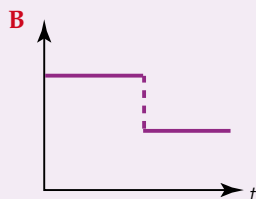
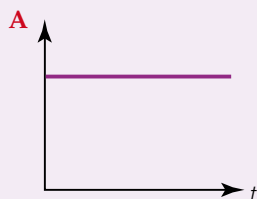
A shopping trolley of mass 5.0 kg, carrying 3.0 kg of potatoes and 2.0 kg of oranges, is given an initial push and moves away with constant horizontal speed of 5.0 m s^{-1} . While it is still moving at this speed, another shopper drops a 10 kg bag of apples vertically into the trolley. Ignore the effects of friction.

7 Determine the subsequent speed of the trolley.

8 Which one of the following statements correctly describes what has happened to the vertical momentum of the apples?

- A It has been converted into the horizontal momentum of the trolley.
- B It has been dissipated as heat and sound.
- C It has been transferred to the Earth.
- D It has been converted into kinetic energy.

9 a Which one of the following graphs best represents the horizontal momentum of the trolley alone as a function of time?



b Which one of these graphs best represents the total horizontal momentum of the trolley and its contents as a function of time?

10 A girl of mass 48 kg runs towards her stationary 2.0 kg skateboard and jumps on. The horizontal velocity of the girl just before she makes contact with the skateboard is 4.0 m s^{-1} to the right. Ignore the effects of friction.

- a Calculate the momentum of the girl just before she lands on the skateboard.
- b Determine the velocity of the girl after she has landed on the skateboard.
- c The girl then travels at this speed for a short time before jumping off, still travelling at the same speed as the skateboard. What is the velocity of the skateboard after she jumps off?

11 A footballer crashes into the fence at the MCG and stops completely. What has happened to his momentum? Has momentum been conserved? Explain.

12 As a high-board diver falls towards a diving pool, her momentum increases. Does this mean that momentum is not conserved? Explain.

2.3

Work, energy and power

Energy

The Universe is made up of *matter* and *radiation*. Matter is the stuff of which things are made—atoms, molecules and subatomic particles. *Energy* is a property of both matter and radiation. When a force acts on a body, causing it to move, the energy of the body changes. If a water droplet absorbs radiation in the form of microwaves, the water molecules will begin to vibrate strongly and the water will gain heat energy. About 100 years ago, Einstein showed that matter and energy were equivalent and that vast amounts of energy could be obtained from small quantities of matter. You might remember from Year 11 how matter is converted into radiation energy during radioactive decay.

Energy comes in many different forms including heat, sound, light, chemical, electrical and nuclear energy. Energy is a *scalar* quantity and is measured in *joules (J)*. In Year 11, you studied a number of different forms of energy. Some of these are outlined below.

Kinetic energy is the energy of *motion*. The Earth has a massive amount of kinetic energy as it moves in its orbit around the Sun. An electron has a minuscule quantity of kinetic energy as it orbits the nucleus of an atom.



The **KINETIC ENERGY** (E_k) of an object is given by:

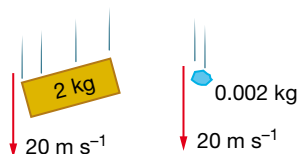
$$E_k = \frac{1}{2}mv^2$$

where E_k = kinetic energy (J)

m = mass (kg)

v = speed (m s^{-1})

(a)



(b)

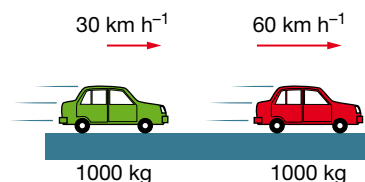


Figure 2.14 (a) The brick and the hailstone are travelling at the same speed, but the brick has 1000 times more kinetic energy because its mass is 1000 times greater. (b) The red car is travelling twice as fast as the green car and has four times as much kinetic energy.

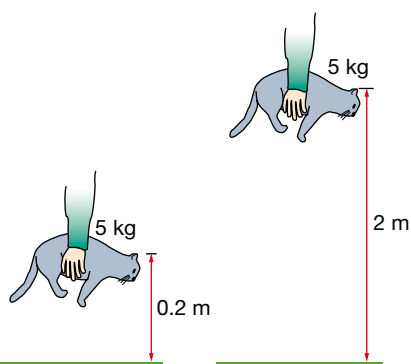


Figure 2.15 When the cat is held 2 m off the ground, it has 10 times more gravitational potential energy than when it is only 0.2 m from the ground.

Gravitational potential energy is energy due to an object's *position* in a gravitational field. This is a form of *stored energy*. A cat that is held at rest above the ground has a small amount of gravitational potential energy. As it is held higher above the ground, the cat's gravitational potential energy increases.



GRAVITATIONAL POTENTIAL ENERGY (U_g) is given by:

$$U_g = mgh$$

where U_g = gravitational potential energy (J)

m = mass (kg)

g = gravitational field strength (N kg^{-1}) ($g = 9.8 \text{ N kg}^{-1}$ near the Earth's surface)

h = height above a reference point (m)

Another form of potential energy that was studied in Year 11 is *elastic* or *strain potential energy*. This will be covered later in this chapter.

Work and energy

The concept of *work* is central to understanding energy. Energy changes occur whenever work is done. A force does work whenever it acts on a body and causes it to undergo a displacement. An unbalanced force does work on a body when it produces a *change in energy*. A weightlifter does work in lifting a barbell, since the applied force results in a displacement, and the gravitational potential energy of the barbell has increased. However, as the weightlifter holds the barbell overhead, no work is being done because the weight is not undergoing a displacement. Work, like energy, is a *scalar* quantity and so does not have a direction associated with it.



WORK (W) is the product of the applied force and the displacement in the direction of the force. When work is done on a body, the energy of the body changes, i.e.

$$W = \Delta E = Fx \cos \theta$$

where W = work (N m, or J)

ΔE = change in energy (J)

F = force (N)

x = displacement (m)

θ = the angle between applied force and direction of motion

From the definition it can be seen that as the angle between the applied force and the direction of motion increases, the amount of work done decreases. When the force acts at 90° to the direction of motion, no work is done at all.

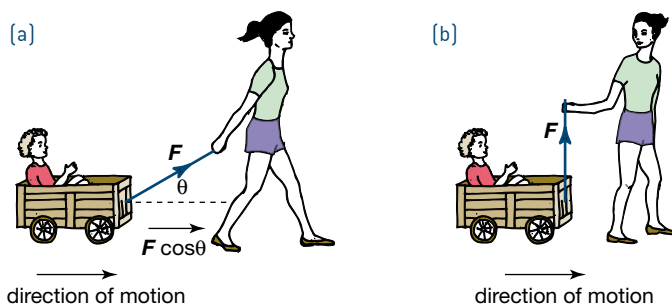


Figure 2.16 (a) If the force acts at an angle to the direction of motion of the cart, the force is less effective in doing work on the cart. The component of the force in the direction of the displacement must be determined. (b) This force does not change the energy of the cart, and so does no work on the cart (provided it does not lift the cart from the ground). There is no component of the pulling force in the direction of motion of the cart.

When information is presented in the form of a force–displacement (F – x) graph, the *work done* and the *change in energy* can be found by calculating the *area* under the graph as shown in Figure 2.17.

Power

Consider the example of two students, Kylie and Nick, each of mass 60 kg, who travel up four flights of stairs. If Kylie sprints up the stairs and Nick walks up at a leisurely pace, then Kylie will have increased her gravitational potential energy at a faster rate than Nick. In other words, Kylie has developed more *power* than Nick.

Physics file

The unit for work is the newton metre (N m), which is called a joule (J) in honour of James Joule, an English physicist who did pioneering work on energy in the 19th century. All forms of energy are measured in joules. The units for gravitational potential energy, for example, can be shown to be equivalent to joules.

$$U_g = mgh$$

$$\text{Units for } U_g = \text{kg m s}^{-2} \text{ m}$$

$$= \text{N m}$$

$$= \text{J}$$

You might like to try this for kinetic energy.

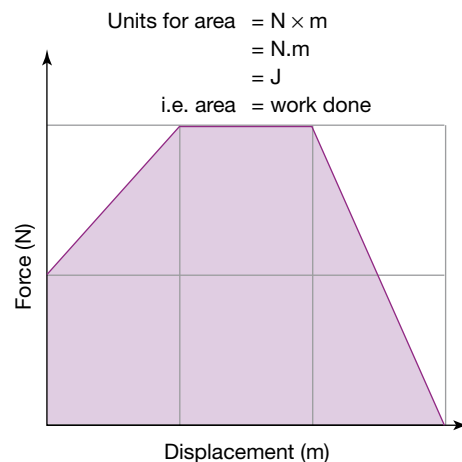


Figure 2.17 The area under a force–displacement (F – x) graph is the work done by the force.

Physics file

The British Imperial unit for power is horsepower (hp); 1 hp = 746 W. This unit dates from the time of the Industrial Revolution, when steam engines began to perform mechanical tasks that had previously been done by horses.



POWER is the rate at which work is done. It is a scalar quantity and is measured in watts [W]; $1 \text{ W} = 1 \text{ J s}^{-1}$.

$$\text{Power} = \frac{\text{work done}}{\text{time taken}}$$

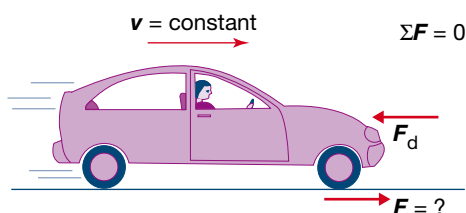


Figure 2.18 The power required to keep an object moving with constant velocity against frictional forces is given by $P = Fv$.

Physics file

All equations for energy originate from the definition of work. If a force \mathbf{F} acts on a body of mass m , causing a horizontal displacement \mathbf{x} , the work done is:

$$W = Fx = max$$

Now $v^2 = u^2 + 2ax$ can be rearranged as:

$$x = \frac{v^2 - u^2}{2a}$$

$$\begin{aligned} \therefore W &= ma \left(\frac{v^2 - u^2}{2a} \right) \\ &= \frac{1}{2}mv^2 - \frac{1}{2}mu^2 \\ &= \Delta E_k \end{aligned}$$

In general, $E_k = \frac{1}{2}mv^2$

Similarly, when a body is lifted at a uniform rate, the lifting force is simply equal to the gravitational force, i.e. mg . If the mass m is lifted through a vertical displacement \mathbf{x} , the work done on the body is:

$$W = Fx = mgx$$

This vertical displacement is the change in height, Δh , in the gravitational field.

Thus:

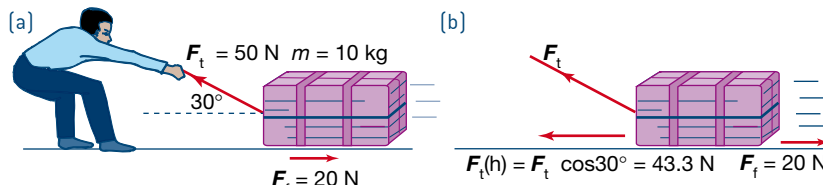
$$W = mg\Delta h = \Delta U_g$$

In general:

$$U_g = mgh$$

Worked example 2.3A

A rope that is at 30.0° to the horizontal is used to pull a 10.0 kg crate across a rough floor. The crate is initially at rest and is dragged for a distance of 4.00 m .



If the tension, F_t , in the rope is 50.0 N and the frictional force opposing the motion is 20.0 N , determine:

- the work done on the crate by the person pulling the rope
- the work done on the crate
- the energy transformed to heat and sound due to the frictional force.

Solution

- a** The work done on the crate by the person is given by:

$$W = Fx \cos \theta = 50.0 \times 4.00 \times \cos 30^\circ = 173 \text{ J}$$

- b** To find the work done on the crate, it is necessary to find the net horizontal force that is acting on the crate.

$$\text{The net horizontal force } \Sigma F_h = 43.3 - 20.0 = 23.3 \text{ N}$$

$$\begin{aligned} \text{Work done} &= \Sigma F_h \times x \\ &= 23.3 \times 4.00 \\ &= 93.2 \text{ J} \end{aligned}$$

This indicates that the crate gains 93.2 J of kinetic energy. As it starts from rest, its final kinetic energy is 93.2 J .

- c** The person pulling has done 173 J of work on the crate, yet the crate has gained only 93.2 J of energy. This indicates that 80 J of energy is transformed to heat and sound energy and so lost to the system due to friction. This could also be determined by finding the work done against friction:

$$\begin{aligned} W &= F_f \times x \\ &= 20.0 \times 4.00 \\ &= 80.0 \text{ J} \end{aligned}$$

Conservation of energy

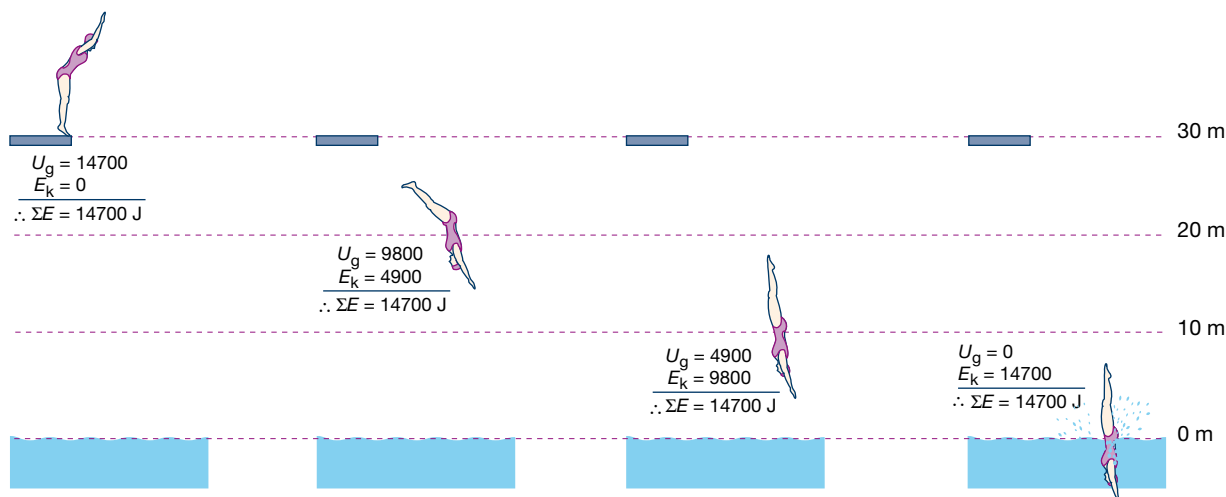
A fundamental principle of nature is that energy is conserved. The law of conservation of energy states that energy is transformed from one form to another, and that the *total amount of energy* in its various forms *remains constant*. In other words, energy cannot be created or destroyed—it can only change from one form to another. For example, when an arrow is fired from a bow, the stored energy in the stretched bow is converted into kinetic energy

as the arrow is released. When the arrow strikes the target, its kinetic energy is converted into heat and sound.

Imagine Sarah, a high-board diver of mass 50 kg diving from a height of 30 m into a pool. When Sarah is on the platform, she has gravitational potential energy of:

$$\begin{aligned}U_g &= mgh \\&= 50 \times 9.8 \times 30 \\&= 14\,700 \text{ J}\end{aligned}$$

The sum of an object's *kinetic and potential energy* is called the *mechanical energy*. Sarah has zero kinetic energy when she is standing on the platform, so her initial mechanical energy is 14 700 J. During the fall, she will lose gravitational potential energy and gain kinetic energy, but the sum of these energies will remain 14 700 J. In other words, her mechanical energy will not change throughout the dive. These energy changes are shown in Figure 2.19.



As she dives towards the water, the gravitational force does work on her. During the first 10 m of the dive, she loses 4900 J of gravitational potential energy and gains an equal amount of kinetic energy. This transfer of gravitational potential to kinetic energy continues until she reaches the water. At this point, Sarah has zero potential energy and 14 700 J of kinetic energy.

Physics file

The world golf authorities have placed restrictions on how energy efficient the face of a golf club can be made. With modern materials and spring-like club faces, manufacturers have been making clubs that are more and more efficient. There is a ruling that the coefficient of restitution (i.e. the energy efficiency) of a club-head must not exceed 0.83. This means that at least 17% of the club-head's mechanical energy must be transformed into heat and sound.

Figure 2.19 The total energy of the diver remains at 14 700 J throughout the dive. She loses gravitational potential energy and gains kinetic energy as she falls.

Worked example 2.3B

A golf ball of mass 110 g is hit straight up into the air at 35 m s^{-1} .

- What is the initial kinetic energy of the golf ball?
- Disregarding drag forces, calculate the height to which the ball travels.

Solution

$$\begin{aligned}\text{a } E_k &= \frac{1}{2}mv^2 \\&= 0.5 \times 0.110 \times (35)^2 \\&= 67 \text{ J}\end{aligned}$$

- The initial potential energy of the ball is zero, so it has 67 J of mechanical energy. Therefore, it has 67 J of mechanical energy when at its maximum height. Here it will be at rest so will have zero kinetic energy. Therefore, at this point the ball will have 67 J of gravitational potential energy.

$$\begin{aligned}U_g &= mgh \\67 &= 0.110 \times 9.80 \times h \\h &= 62 \text{ m}\end{aligned}$$

If air resistance is ignored, the ball will reach a height of 62 m.



Figure 2.20 As bungee jumpers fall, gravitational potential energy is transformed first into kinetic energy and then into elastic potential energy in the stretched bungee rope.

Elastic and inelastic collisions

In all collisions, momentum is conserved. Energy, in all its forms, is also conserved every time. These two principles—conservation of momentum and conservation of energy—hold true at all levels: at the grand scale of stars and galaxies, in everyday situations such as collisions between balls on a billiard table, through to the atomic interactions between the particles that make up matter.

Although energy is conserved in a collision, it is unusual for a particular form of energy to be conserved. However, there are situations in which no kinetic energy is lost; these are elastic collisions. Perfectly elastic collisions do not exist in everyday situations, but they do exist in the interactions between atoms and subatomic particles. A collision between two billiard balls is almost elastic because very little of their kinetic energy is transformed into heat and sound energy.



An **ELASTIC COLLISION** is one in which both the kinetic energy and momentum of the system are conserved.

Inelastic collisions can vary from almost elastic to perfectly inelastic. *Almost elastic* collisions include those where little friction acts, for example between billiard balls and between air track gliders with repelling magnets. Collisions such as a bouncing basketball, a gymnast on a trampoline and a tennis ball being hit are moderately elastic with about half the kinetic energy of the system being retained. *Perfectly inelastic* collisions are those in which the colliding bodies stick together after impact. Some car crashes, a collision between a meteorite and the Moon, and a collision involving two balls of plasticine, would be perfectly inelastic. In these collisions, much, and sometimes all, of the initial kinetic energy of the system is lost.



An **INELASTIC COLLISION** is a collision in which momentum is conserved but kinetic energy is transformed into other forms of energy.

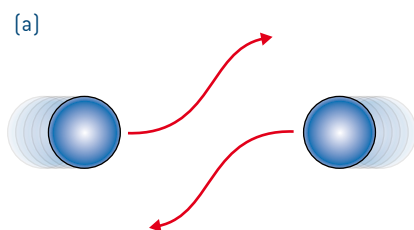


Figure 2.21 (a) When two electrons collide, the force of repulsion due to their like charges prevents them from coming into contact. No kinetic energy is lost in this collision; it is perfectly elastic. (b) During most collisions, kinetic energy is transformed into heat and sound and is so lost to the system of objects. These are inelastic collisions.

Worked example 2.3C

A car of mass 1000 kg travelling west at 20 m s^{-1} crashes into the rear of a stationary bus of mass 5000 kg. The vehicles lock together on impact.

- What is their joint velocity immediately after the collision?
- What is the total kinetic energy of the system before the collision?
- What is the total kinetic energy of the system after the collision?
- Is this an elastic or inelastic collision? Explain.

Solution

- a** For this problem, west will be taken as the positive direction. Conservation of momentum must be used to determine the final velocity, \mathbf{v}' , of the vehicles.

$$\Sigma \mathbf{p}_i = \Sigma \mathbf{p}_f$$

$$\mathbf{p}_i(\text{car}) + \mathbf{p}_i(\text{bus}) = \mathbf{p}_f(\text{bus and car})$$

$$(1000 \times 20) + 0 = (1000 + 5000)\mathbf{v}'$$

$$\mathbf{v}' = \frac{20000}{6000} = 3.3 \text{ m s}^{-1}, \text{ i.e. } 3.3 \text{ m s}^{-1} \text{ west}$$

- b** The initial kinetic energy of the system is the initial kinetic energy of the car. It has a mass of 1000 kg and is moving at 20 m s^{-1} prior to the collision.

$$E_k = \frac{1}{2}mv^2 = 0.5 \times 1000 \times 20^2 = 2.0 \times 10^5 \text{ J}$$

- c** After the collision, the car and bus are locked together and moving at 3.3 m s^{-1} . Their total kinetic energy is:

$$E_k = \frac{1}{2}mv^2 = 0.5 \times (1000 + 5000) \times 3.3^2 = 3.3 \times 10^4 \text{ J}$$

- d** This is an inelastic collision. A large amount of kinetic energy has been removed from the system. This kinetic energy has been transformed into heat and sound and permanently deforming the materials.



2.3 summary

Work, energy and power

- Energy is the ability of an object to do work.
- Work, W (measured in joules, J), is the product of a force and its displacement in the direction of the force:
$$W = Fx \cos \theta$$
- Energy and work are scalar quantities.
- When an unbalanced force does work on an object, the energy of the object changes.
- Power, P (measured in watts, W), is the rate at which work is done:
$$P = \frac{\text{work done}}{\text{time taken}} = \frac{W}{t} = \frac{Fx}{t} = Fv$$
- Kinetic energy is the energy of motion and is given by:
$$E_k = \frac{1}{2}mv^2$$
- Gravitational potential energy is energy due to an object's position in a gravitational field and is given by:
$$U_g = mgh$$
- Mechanical energy is the sum of an object's kinetic and potential energies.
- Whenever energy is transformed from one form to another, the total amount of energy remains constant. This conservation of energy is a fundamental principle in nature.
- An elastic collision is where the kinetic energy is conserved throughout the collision.
- An inelastic collision is one where some kinetic energy is transformed into heat and sound. Collisions are usually inelastic.



2.3 questions

Work, energy and power

For the following questions, assume that g is 9.8 m s^{-2} , and that the effects of air resistance are negligible.

- 1** A crane on a building site lifts an 800 kg load from ground level to a vertical height of 90 m at a

constant speed of 2.0 m s^{-1} . Ignore the mass of the cable.

- a** How much work is done by the crane in lifting the load through this distance?



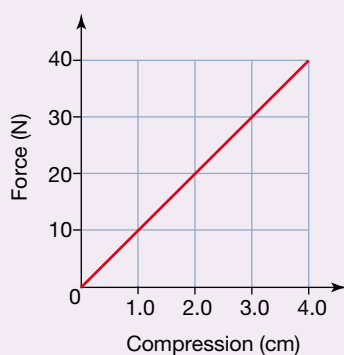
- b** What is the mechanical energy of the load when it has reached a vertical height of 50 m?
- c** Calculate the power developed by the crane motor during this operation.

The following information applies to questions 2–5.

A child uses a leash to drag a reluctant 2.0 kg puppy, across a floor. The leash is held at an angle of 60° to the horizontal and applies a force of 30 N on the puppy, which is initially at rest. A constant frictional force of 10 N acts on the dog as it is dragged for a distance of 2.5 m.

- 2** For the 2.5 m distance, calculate the work done by:
 - a** the horizontal component of the 30 N force
 - b** the frictional force
 - c** the resultant horizontal force.
- 3** **a** Calculate the change in kinetic energy of the puppy as it is dragged along.
b How much energy has dissipated as heat during this time?
c Which force is responsible for some energy being transformed into heat energy?
- 4** How much work does the vertical component of the 30 N pulling force do on the dog during its motion? Justify your answer.
- 5** How long does the pup take to travel the 2.5 m?
- 6** A 200 g snooker ball with initial velocity 9.0 m s^{-1} to the right collides with a stationary snooker ball of mass 100 g. After the collision, both balls are moving to the right. The 200 g ball has a speed of 3.0 m s^{-1} , while the 100 g ball has a speed of 12 m s^{-1} .
 - a** Calculate the total kinetic energy of the system before the collision.
 - b** Determine the total kinetic energy of the system after the collision.
 - c** Is this collision elastic or inelastic? Justify your answer.
 - d** Is this situation realistic? Justify your answer.

The following information applies to questions 7 and 8.



A gymnast decides to improve the strength of her grip by repeatedly squeezing a small rubber ball of mass 100 g whose force–compression graph is shown.

- 7** **a** How much work must the gymnast do to compress the ball by 10 mm?
b Discuss the energy changes that occur in the ball during one cycle of compression and release.
c How much power is expended by the gymnast if she performs 120 identical compressions of 10 mm in a 60 s interval? (Assume that no work is done on release.)
- 8** After the gymnast finishes her exercises, she throws the ball at a solid brick wall. The ball strikes the wall horizontally with a speed of 4.0 m s^{-1} . Calculate the maximum compression produced in the ball during this impact.
- 9** A firefighter of mass 80 kg slides down a 5.0 m pole, and by gripping the pole as he slides down, has an acceleration of 5.0 m s^{-2} . His velocity at the top of the pole is zero.
 - a** Calculate the change in the firefighter's gravitational potential energy during the slide down the pole.
 - b** What was the corresponding change in kinetic energy?
 - c** Explain why the change in the firefighter's gravitational potential energy was not equal in magnitude to the change in his kinetic energy.
 - d** Calculate the work done on this firefighter by the gravitational field of the Earth.
 - e** How much heat energy was produced as the firefighter slid down the pole?
- 10** Two identical bowling balls, each of mass 4.0 kg, move towards each other across a frictionless horizontal surface with equal speeds of 3.0 m s^{-1} . During the collision, 20 J of the initial kinetic energy is transformed into heat and sound. After the collision, the balls move in opposite directions away from each other.
 - a** Is momentum conserved in this collision?
 - b** Is this an elastic or inelastic collision?
 - c** Calculate the speed of each ball after the collision.

2.4 Hooke's law and elastic potential energy

Hooke's law

Springs and other elastic materials are very useful in everyday life. The keys on computer keyboards are sometimes mounted on small springs. Springs are also used in car suspensions, trampolines, wind-up toys, watches and door-closing mechanisms. Running shoes, tennis balls and racquets, archery bows and bungee ropes employ elastic materials to help them perform their tasks.

In order to determine the effect of doing work on a spring, we need to examine the behaviour of a spring under different forces. Consider a simple experiment where a cart is attached to a spring that is progressively stretched (Figure 2.22). As the spring is stretched more, it exerts a larger force on the cart.

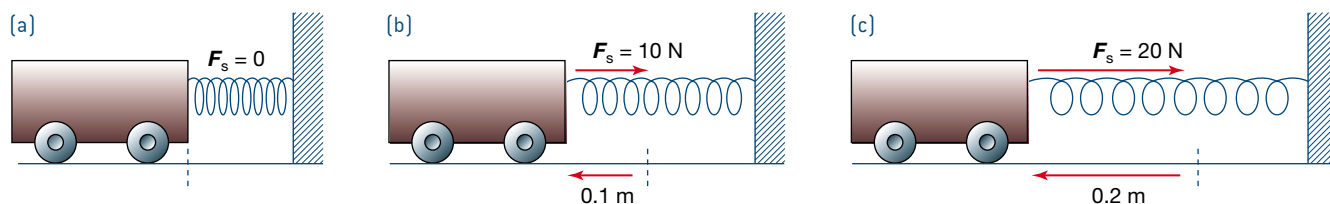


Figure 2.22 The force exerted by the spring increases as it is stretched. (a) The spring is unstretched, i.e. at its natural length, and so exerts zero force on the cart. (b) As the spring is stretched to the left, it exerts a force to the right on the cart. This is called a restoring force. (c) The extension of the spring has doubled, and the force that it exerts has also doubled. This spring obeys Hooke's law.

For a light spring, there is a direct relationship between the force exerted by the spring and the extension in the spring. The force required to double the extension of a spring is itself double. In other words, there is a direct relationship between F and x for the spring. The relationship between the force and the extension is known as *Hooke's law*.



HOOKE'S LAW states that the force exerted by a spring is directly proportional to, but opposite in direction to, the spring's extension or compression:

$$F_s = -kx$$

where F_s = force (N)

k = force constant (N m^{-1})

x = amount of extension or compression (m)

The *spring constant*, k (also called a force constant), indicates the *stiffness* of the material. Its units are newtons per metre (N m^{-1}). A spring constant of 25 N m^{-1} indicates that for every metre that the spring is stretched or compressed, a force of 25 N is required. This does not necessarily mean that the spring can be stretched by 1 m, but it tells us that the force and the change in length are in this proportion. The spring constant can be determined as the *gradient* of the F - x graph.

The force constant of the spring in Figure 2.23 is therefore:

$$\begin{aligned} k &= \text{gradient} \\ &= \frac{20}{0.2} \\ &= 100 \text{ N m}^{-1} \end{aligned}$$



PRACTICAL ACTIVITY 11

Conservation of energy in springs



PRACTICAL ACTIVITY 12

Hooke's law

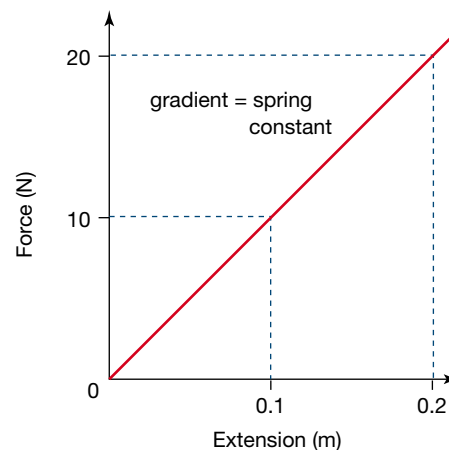


Figure 2.23 The F - x graph for this spring is a straight line passing through the origin. For this reason, it is said to obey Hooke's law and is known as an *ideal* spring.

Physics file

The ropes used by rock climbers have elastic properties that can save lives during climbing accidents. Ropes that were used in the 19th century were made of hemp, which is strong but does not stretch a lot. When climbers using these ropes fell, they stopped very abruptly. The resulting large forces acting on the climbers caused many serious injuries. Modern ropes are made of a continuous-drawn nylon fibre core and a protective textile covering. They have a slightly lower spring constant and stretch significantly (up to several metres) when stopping a falling climber. This reduces the stopping force acting on the climber. Ropes with even lower spring constants are suitable for bungee jumping. Rock climbers tend to avoid these ropes—bouncing up and down the rock face is not advisable!

Physics file

The relationship for the elastic potential energy of materials that obey Hooke's law is derived as follows:

$$U_s = \text{area under an } F\text{-}x \text{ graph} = \frac{1}{2}F \times x$$

However, according to Hooke's law

$$F = kx, \text{ so:}$$

$$U_s = \frac{1}{2}kx^2$$

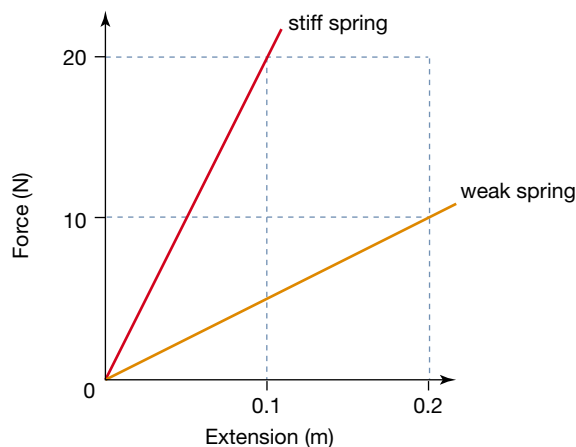


Figure 2.24 Both springs represented in this graph are ideal. They obey Hooke's law, but they have different degrees of stiffness. The stiff spring has a force constant of 200 N m^{-1} , i.e. it takes a force of 200 N to stretch it by 1 m. The force constant of the weak spring is just 50 N m^{-1} . The stiffer spring has the steeper line on the $F\text{-}x$ graph.

Any material or object in which there is a direct relationship between the amount of extension or compression and the force that it exerts is said to follow Hooke's law. The graph of its behaviour would show a straight line passing through the origin (Figure 2.23). However, if the force applied to the spring is very large, the spring becomes permanently distorted and will no longer obey Hooke's law (Figure 2.25). The point at which the graph deviates from the straight line (i.e. no longer follows Hooke's law) is called the *elastic limit* for the spring.

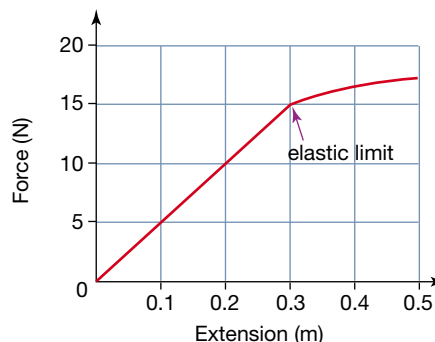


Figure 2.25 When a spring is over-stretched so that it becomes permanently deformed and does not return to its original length, it is said to have exceeded its elastic limit, i.e. it no longer obeys Hooke's law. The elastic limit of this spring is 15 N.

Elastic potential energy

Any spring that has been stretched or compressed has stored *elastic potential energy*. This means that the spring is able to do work on another object by exerting a force over some distance as the spring resumes its original length. The energy stored in a spring is also called *strain potential energy*. As discussed in section 2.3, the work done by a constant force F is given by $W = Fx$, but this rule cannot be used to find the work done by springs, because the force that acts is not constant. The force becomes larger as the extension increases. In this event, the *area* under a force–displacement ($F\text{-}x$) graph is the *work done* on an object or the *energy change* that has occurred; we can use this approach to find the work done by a spring.



ELASTIC POTENTIAL ENERGY (U_s) can be determined by finding the area under an F - x graph.

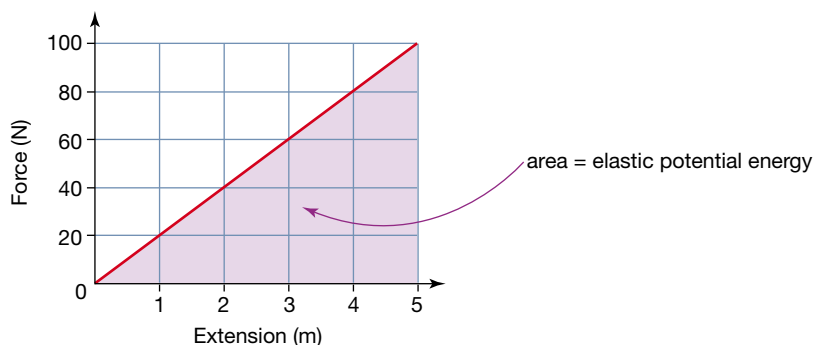


Figure 2.26 The energy stored in a spring is given by the area under the F - x graph.



ELASTIC POTENTIAL ENERGY for an ideal spring is given by:

$$U_s = \frac{1}{2}kx^2$$

where U_s = elastic potential energy (J)

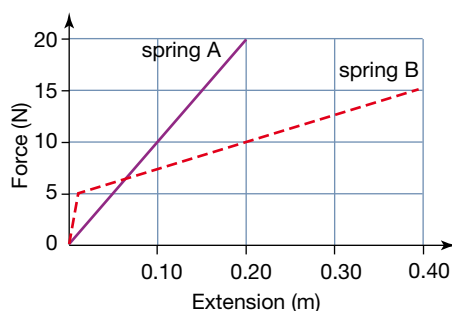
k = force constant (N m^{-1})

x = compression or extension (m)

Worked example 2.4A

The characteristics of two springs A and B are shown in the graph. Spring B does not follow Hooke's law.

- Discuss the stiffness of each spring.
- How much work has been done to stretch spring A by 20 cm?



Solution

- The stiffness of each spring is indicated by the spring constant, which is given by the gradient of each line. Spring A has a spring constant of 100 N m^{-1} . The stiffness of spring B changes. It is very stiff when the force acting on it is less than 5 N. Then it has a spring constant of 25 N m^{-1} when the force exceeds 5 N.
- The strain energy in spring A when it is extended by 0.20 m can be found in two ways. Since spring A obeys Hooke's law, its elastic potential energy is given by:

$$\begin{aligned} U_s &= \frac{1}{2}kx^2 \\ &= 0.5 \times 100 \times 0.20^2 \\ &= 2.0 \text{ J} \end{aligned}$$

Its strain potential energy can also be found by calculating the area under the graph:

$$\begin{aligned} \text{Area} &= 0.5 \times 0.20 \times 20 \\ &= 2.0 \text{ J} \end{aligned}$$

Physics file

A car tyre originally developed by Michelin and Goodyear is claimed to reduce fuel consumption by an average of 5%. It is made using silica, an elastic material, in the tread. These tyres have a resistance to motion that is 20% lower than that of conventional tyres. This leads to improved fuel efficiency because the tyres heat up less and so waste less energy. A tyre has a resistance to motion, or rolling resistance, because it continually flattens out as it comes into contact with the road and then returns to its normal shape after contact has ceased.



Figure 2.27 Tyres deform as they rotate. The continual and rapid deformation that takes place in the tyre material is responsible for the tyre heating up. This heating effect reduces the fuel efficiency of the car. Recently developed tyres use silica in the tread to greatly reduce this heating.

Worked example 2.4B

A toy plane is launched by using a stretched rubber band to fire the plane into the air. The rubber band is stretched by 25 cm and has a spring constant of 120 N m^{-1} . The mass of the plane is 160 g. Assume that the rubber band follows Hooke's law and ignore its mass when answering these questions.

- How much potential energy is stored in the rubber band at this extension?
- Calculate the speed with which the model plane is launched.
- Without using calculations, describe how the kinetic energy of the plane will compare if the rubber band is stretched by twice as much when launching. (Assume that the rubber band continues to behave ideally.)

Solution

- $$U_s = \frac{1}{2}kx^2$$

$$= 0.5 \times 120 \times (0.25)^2$$

$$= 3.8 \text{ J}$$
- Assuming conservation of mechanical energy, the plane will have 3.8 J of kinetic energy as it is released, so:

$$\frac{1}{2}mv^2 = 3.8$$

$$v = 6.8 \text{ m s}^{-1}$$

The launch speed of the plane is 6.8 m s^{-1} .
- If the extension of the rubber band is doubled, the elastic energy stored within it will quadruple. This is evident from $U_s = \frac{1}{2}kx^2$ which indicates a square relationship between potential energy and extension. Since there is now four times as much strain energy, the plane will have four times as much kinetic energy when it is released and so be launched with double the speed.

Physics in action

Energy changes in pole vaulting

The world record for the men's pole vault is over 6 m—about as high as a single-storey house! The women's record is just over 5 m. During the jump, a number of energy transformations take place. The athlete has kinetic energy as she runs in. This kinetic energy is used to bend the pole and carry the athlete forwards over the bar. As the pole bends, energy is stored as elastic potential energy. The athlete uses this stored energy to increase her gravitational potential energy and, hopefully, raise her centre of mass over the bar. Once the pole has been released and the bar has been cleared, the gravitational potential energy of the athlete is transformed into kinetic energy as she falls towards the mat.

We can analyse the energy changes by making some assumptions about the athlete and the jump. Let us say that the athlete has a mass of 60 kg and runs in at 7.0 m s^{-1} . We will treat the athlete as a point mass located at her centre of mass, 1.2 m above the ground. The athlete raises her centre of mass to a height of 5.0 m as she clears the bar, and her speed at this point is just 1.0 m s^{-1} . As she plants the pole in the stop, the pole has not yet been bent and so it has no elastic potential energy. Using:

$$\Sigma E = E_k + U_g = \frac{1}{2}mv^2 + mgh$$

we calculate the vaulter's total energy at this point to be 2180 J.

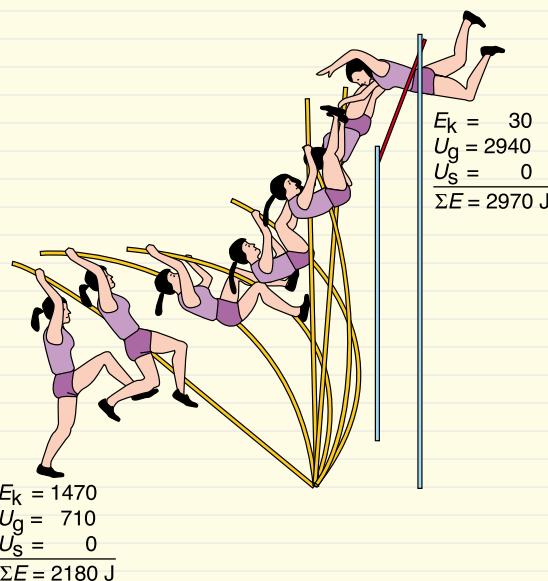


Figure 2.28 These diagrams, drawn at equal time intervals, indicate that this vaulter slows down as she nears the bar. Her initial kinetic energy is stored as elastic potential energy in the bent pole, and finally transformed into gravitational potential energy and kinetic energy, enabling her to clear the bar.

When the vaulter passes over the bar, the pole is straight again and so has no elastic potential energy. Taking the ground as zero height, and using the same relationship as above, we find that the vaulter's total energy is now 2970 J. This does not seem consistent with the conservation of energy. Where has the extra 790 J come from? The answer is, from the muscles in her body.

Just before the athlete plants the pole, she raises it over her head. Then after the pole is planted, but before she leaves the ground, the athlete uses her arms to bend the pole (Figure 2.29). She pulls downwards on the pole with one arm while the other arm pushes upwards. The effect of these forces is to do work on the pole and store some extra elastic potential energy in it. This work will be converted into gravitational potential energy later in the jump. Energy has also been put into the system by the muscles of the athlete as they do work after she has left the ground. Throughout the jump, she has used her arm muscles to raise her body higher. At the end of the jump, she is actually ahead of the pole and pushing herself up off it. In effect, she has been pushing off the ground by using the pole.

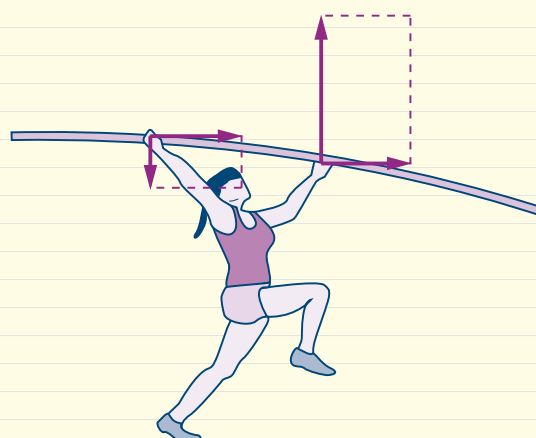


Figure 2.29 As the pole is planted, the vaulter uses her arms to bend the bar. The forces are shown by the vectors. By bending the bar, the athlete has stored energy which will later be transformed into gravitational potential energy.



2.4 summary

Hooke's law and elastic potential energy

- The relationship between the force, F , exerted by a spring and its extension or compression, x , is described by Hooke's law:

$$F = -kx$$

- k is the spring constant or force constant, and is measured in N m^{-1} . The value of the spring constant is given by the gradient of the F - x graph and is an indication of the stiffness of the spring.

- The energy stored in an elastic material that has been stretched or compressed is elastic potential energy or strain energy, measured in joules.
- The elastic potential energy stored in any material can be determined by finding the area under the F - x graph.
- The amount of elastic potential energy, U_s , stored in an ideal elastic material (i.e. one that follows Hooke's law) can also be found by:

$$U_s = \frac{1}{2}kx^2$$

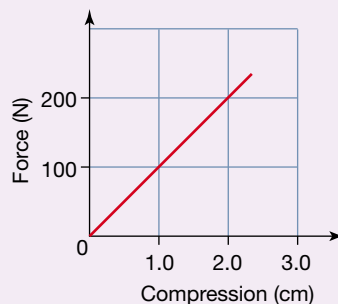


2.4 questions

Hooke's law and elastic potential energy

In the following questions, assume that $g = 9.8 \text{ m s}^{-2}$.

- Two different springs S_1 and S_2 have the force constants: $S_1: k = 250 \text{ N m}^{-1}$, $S_2: k = 1000 \text{ N m}^{-1}$. Which of the following is the best explanation why spring S_2 has a larger force constant than spring S_1 ?
A Spring S_2 is longer than spring S_1 .
B Spring S_2 is shorter than spring S_1 .
C Spring S_2 is made from a different material to spring S_1 .
- The force-compression graph for the spring from an old pinball machine is shown.

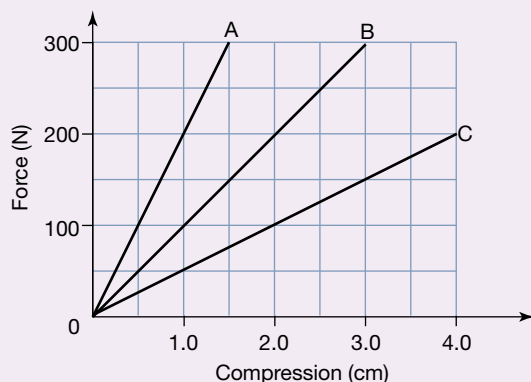


- Determine the magnitude of the force needed to compress this spring by 2.0 cm.



- b** Calculate the spring constant for this spring.
- c** How much work must be done in order to compress this spring by 2.0 cm?
- d** Calculate the elastic potential energy stored in the spring when it is compressed by 1.0 cm.
- e** Plot a graph of stored energy versus compression for this spring.

The following information applies to questions 3 and 4. The diagram shows the force–compression graph for three different springs: A, B and C.



- 3 a** Calculate the force constant for each spring.
- b** Rank these springs in order of increasing stiffness.
- c** Calculate the strain energy stored in each spring at a compression of 1.0 cm.
- 4** Springs A and B are compressed so that they hold equal amounts of elastic potential energy. Calculate the value of the ratio of the compression of A to that of B.

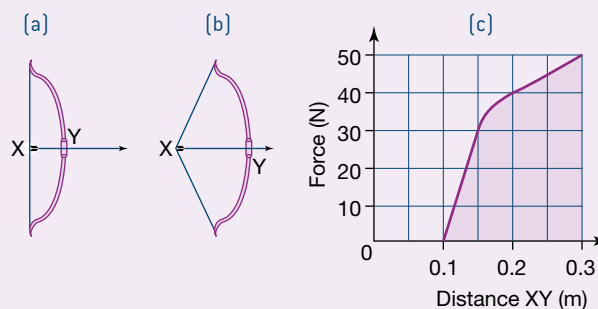
The following information applies to questions 5 and 6. A light spring with an unextended length of 30 cm is hung vertically from a fixed support. A student attaches a 75 g mass to the end of the spring, causing it to stretch by 5.0 cm.

- 5 a** Calculate the spring constant of the spring.
- b** How much strain potential energy is stored in the spring at this extension?
- 6** The 75 g mass is removed and a new object is hung from the spring. While supporting this new object, the spring length increases to 45 cm. Assuming that the spring obeys Hooke's law, calculate:
 - a** the strain energy stored in the spring while it is supporting this mass
 - b** the mass of the new object.

The following information applies to questions 7 and 8. A dynamics cart has a total mass of 2.5 kg and is travelling at 1.5 m s^{-1} towards a wall. The spring bumper of the cart is extended. It has a spring constant of 400 N m^{-1} . As the cart collides with the wall, the spring bumper is compressed and the cart momentarily stops. Ignore the mass of the spring in your calculations.

- 7 a** What is the kinetic energy of the cart before impact?
- b** How much strain energy is stored in the spring as the cart is brought to rest?
- 8 a** By how much is the spring compressed when the cart is brought to rest?
- b** How much kinetic energy does the cart have when the spring is compressed by 4.0 cm?

The following information applies to questions 9 and 10. An Australian archer purchased a new bow for the Beijing Olympics. Diagram (a) shows the bow with the string in the unextended position. When the bow is loaded, the distance XY increases, as shown in (b). Increasing the distance XY increases the force that the string exerts on the arrow. Diagram (c) indicates the force exerted on the arrow against XY as the distance XY is increased from 0.10 m to 0.30 m.



- 9 a** How much work does the archer do on the bow in increasing XY from 0.10 m to 0.30 m?
- b** Explain what happens to this energy as the archer is extending the string.
- 10** When XY is 0.30 m, the archer releases the string.
 - a** How much work is done on the arrow as the string returns to its original position?
 - b** What assumption have you made in your calculation?

2.5 Circular motion

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 centres of their galaxies. In the next chapter, we will be examining planetary and satellite motion in detail. This section serves as an explanation of the nature of circular motion.

Uniform circular motion

An athlete in a hammer throw event is swinging the ball in a horizontal circle with a constant speed of 25 m s^{-1} (Figure 2.31). As the hammer travels in its circular path, its *speed is constant* but its *velocity is continually changing*. At this point, it is important to remember that velocity is a vector. Since the direction of the hammer is changing, so too is its velocity—even though its speed is constant. The velocity of the hammer at any instant is tangential to its path. At one instant, the hammer is travelling at 25 m s^{-1} north, then an instant later at 25 m s^{-1} west, then 25 m s^{-1} south, and so on.

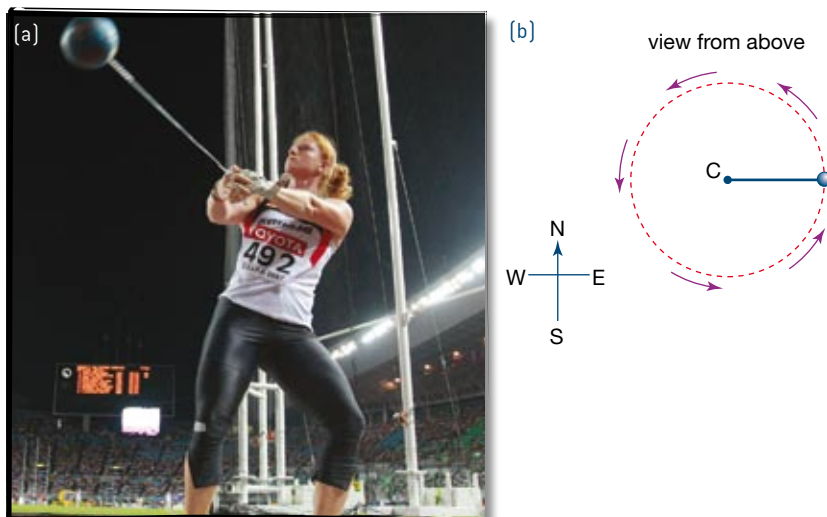


Figure 2.31 (a) The hammer, a ball of steel of mass 7.26 kg for men and 4.0 kg for women, is swung in a circular path before being released. The world record distance for a hammer throw is almost 90 m for men and almost 80 m for women. (b) The velocity of any object moving in a uniform circular path is continually changing, even though its speed remains constant.

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.



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

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

where $v = \text{speed (m s}^{-1}\text{)}$

$r = \text{radius of circle (m)}$

$T = \text{period of motion (s)}$

Physics file

This wind generator is part of a wind farm at Codrington in south-west Victoria. The tower is 50 m high, as high as one of the MCG 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 2.30 The tips of these wind-generator blades are travelling in circular paths at speeds of over 200 km h^{-1} .

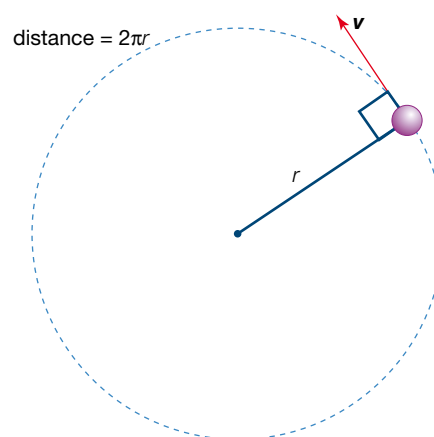


Figure 2.32 The average speed of an object moving in a circular path is given by the distance travelled in one revolution [the circumference] divided by the time taken [the period, T].

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 which track is being played. When the inner track is being played, the disk rotates at about 500 revolutions per minute (rpm). This is 8.3 revolutions each second.



PRACTICAL ACTIVITY 13

Centripetal force

Worked example 2.5A

An athlete is swinging a hammer of mass 7.0 kg in a circular path of radius 1.5 m. Calculate the speed of the hammer if it completes 3.0 revolutions per second.

Solution

The period of the hammer is: $T = \frac{1}{f} = \frac{1}{3.0} = 0.33 \text{ s}$

The speed is: $v = \frac{2\pi r}{T} = \frac{2 \times \pi \times 1.5}{0.33} = 29 \text{ m s}^{-1}$

Worked example 2.5B

A compact disk is rotating at a rate of 8.3 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 8.3 revolutions per second, its period is:

$T = \frac{1}{f} = \frac{1}{8.3} = 0.12 \text{ s}$

The speed of a point on the inner track is:

$v = \frac{2\pi r}{T} = \frac{2 \times \pi \times 2.25}{0.12} = 120 \text{ cm s}^{-1}$ or 1.2 m s^{-1}

At the outer rim of the disk, it has slowed to around 200 rpm or 3.3 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 2.33a, 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 2.33b).

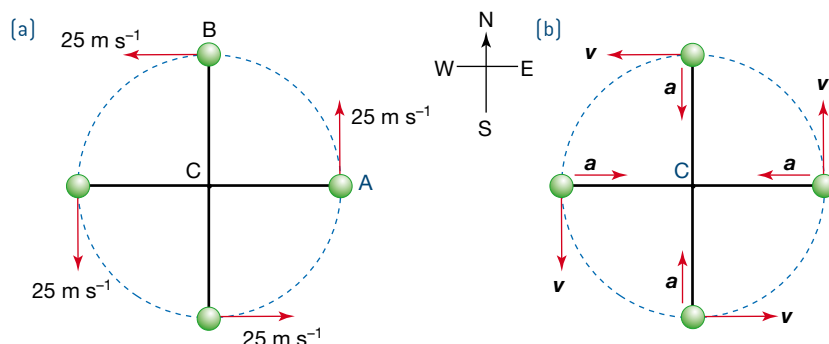


Figure 2.33 (a) The velocity of the hammer at any instant is tangential to its path. (b) A body moving in a circular path has an acceleration towards the centre of the circle. This is known as a centripetal acceleration.

The centripetal acceleration of an object moving in a circular path can be found from the relationship:

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

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

$$v = \frac{2\pi r}{T}, \text{ so } a = \frac{v^2}{r} = \left(\frac{2\pi r}{T}\right)^2 \times \frac{1}{r} = \frac{4\pi^2 r}{T^2}$$



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

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

where a = centripetal acceleration (m s^{-2})

v = speed (m s^{-1})

r = radius of circle (m)

T = period of motion (s)

Forces that cause circular motion

As with all forms of motion, an analysis of the forces that act 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 2.34).

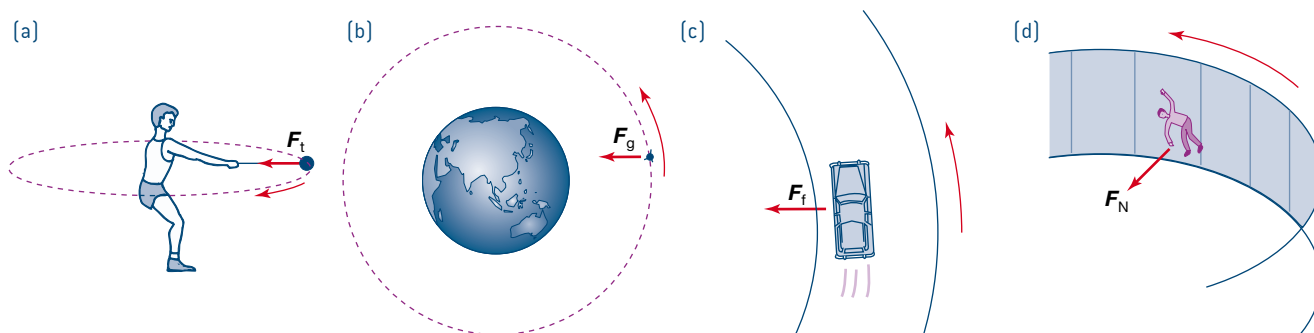


Figure 2.34 The centripetal force that produces a centripetal acceleration and hence a circular motion is provided by different real forces. (a) In a hammer throw or for any other object rotated while attached to an arm or wire, it is the tension in the arm or wire that provides the centripetal force. (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 2.35), 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:

$$\Sigma F = ma = \frac{mv^2}{r} = \frac{4\pi^2 rm}{T^2} \text{ towards the end of the circle}$$

where ΣF = net or resultant force on object (N)

m = mass (kg)

v = speed (m s^{-1})

r = radius of circle (m)

T = period of motion (s)

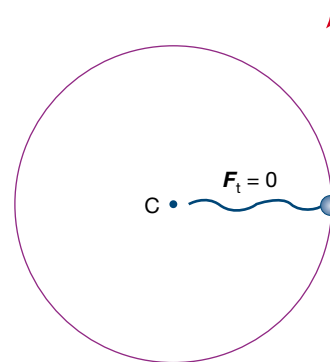


Figure 2.35 When a hammer ball is released by the thrower, it will travel in a straight line at a tangent to its circular path.

Physics file

People moving in circular paths often mistakenly think that there is an outward 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 (the normal force) that is holding them in 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 2.36).



Figure 2.36 There is a large force from the wall (a normal force) that causes these people to travel in a 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.

Worked example 2.5C

An athlete in a hammer throw event is swinging the ball of mass 7.0 kg in a horizontal circular path. Calculate the tension in the wire if the ball is:

- moving at 20 m s^{-1} in a circle of radius 1.6 m
- moving at 25 m s^{-1} in a circle of radius 1.2 m.

Solution

- The centripetal acceleration is:

$$\begin{aligned} a &= \frac{v^2}{r} \\ &= \frac{20^2}{1.6} \\ &= 250 \text{ m s}^{-2} \text{ towards the centre} \end{aligned}$$

The tension is producing circular motion:

$$\begin{aligned} F_t &= \Sigma F \\ &= ma \\ &= 7.0 \times 250 \\ &= 1.8 \times 10^3 \text{ N towards the centre} \end{aligned}$$

- $a = \frac{v^2}{r}$
 $= \frac{25^2}{1.2} = 520 \text{ m s}^{-2} \text{ towards the centre}$

$$\begin{aligned} F_t &= \Sigma F \\ &= ma \\ &= 7.0 \times 520 \\ &= 3.6 \times 10^3 \text{ N towards the centre} \end{aligned}$$

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, $W = Fx \cos \theta$, but the force is directed at 90° to the instantaneous displacement, so $\cos \theta = 0$. 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.



2.5 summary

Circular motion

- 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:

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

- 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:

$$\Sigma F = ma = \frac{mv^2}{r} = \frac{4\pi^2 m r}{T^2}$$

- 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.



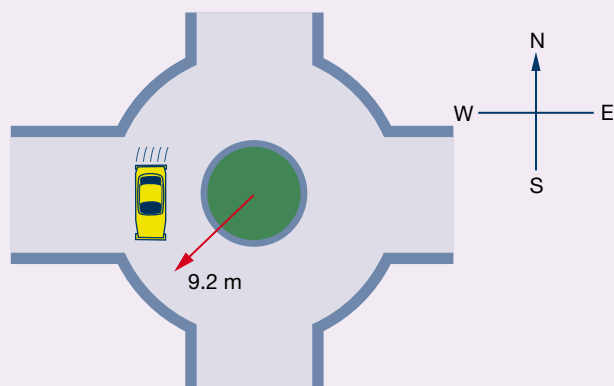
2.5 questions

Circular motion

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 1200 kg is travelling on a roundabout in a circular path of radius 9.2 m . The car moves with a constant speed of 8.0 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:
 - a what is the speed of the car?
 - b what is the velocity of the car?
 - c what is the 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 kg is skating in a horizontal circle of radius 1.5 m at a constant speed of 2.0 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 2.5 kg hammer in a horizontal circle of radius 0.80 m at 2.0 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.

The following information applies to questions 8–10.

Frank and Col are flying their remote-controlled model plane. It has a mass of 1.6 kg and travels in a horizontal circular path of radius 62 m with a speed of 50 km h^{-1} . The plane is controlled by a radio transmitter so there are no strings attached.

- 8 Calculate the period of its motion.
- 9 Determine the magnitude of the net force that is acting on the plane.
- 10 Discuss the nature of the force that is enabling the plane to move in a circular path.

2.6

Aspects of horizontal circular motion

In the previous section, we discussed relatively simple situations involving uniform circular motion in a horizontal plane. Now we will examine some more complex situations involving this type of motion. First, you may have played Totem Tennis at one time. Here the ball is attached to a string and can travel in a horizontal circle, although the string itself is not horizontal. Second, you might have been to a racing track like the Calder Thunderdome where cars travel in circular paths at high speed. The track has banked corners that enable the racing cars to travel at much higher speeds. Finally, you might have watched Casey Stoner racing his motorbike around the corners at Phillip Island and wondered why he leaned his bike over almost to the ground as he turned.

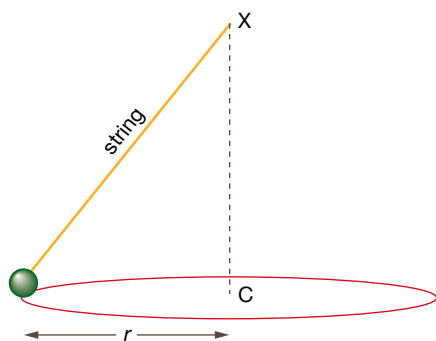


Figure 2.37 This ball is travelling in a horizontal circular path of radius r . The centre of its circular motion is at C.

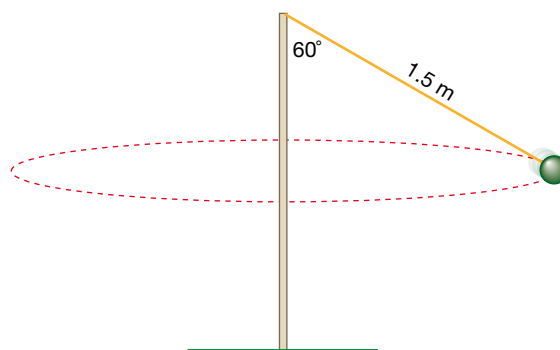
Ball on a string

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 2.37. 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.

Worked example 2.6A

During a game of Totem Tennis, the ball of mass 150 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.

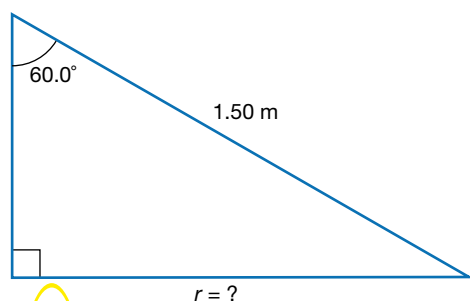


- Calculate the radius of the ball's circular path.
- Draw and identify the forces that are acting on the ball at the instant shown in the diagram.
- Determine the net force that is acting on the ball at this time.
- Calculate the size of the tensile force in the cord.
- How fast is the ball travelling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

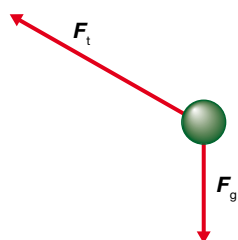
Solution

- The centre of the circular path is not the top end of the cord, but is where the pole is level with the ball. Trigonometry and a distance triangle can be used to work this out.

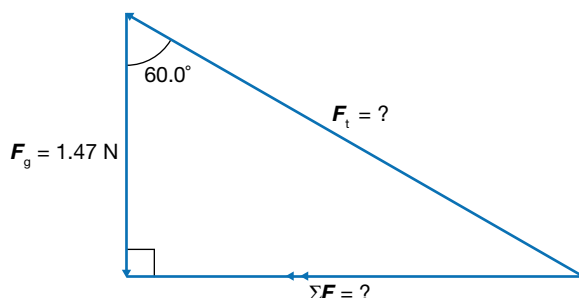
$$r = 1.50 \sin 60.0^\circ = 1.30 \text{ m}$$



- b** There are two forces acting—the tension in the cord, F_t , and gravity, F_g . These forces are unbalanced.



- c** The ball has an acceleration that is towards the centre of its circular path. This is horizontal and towards the left at this instant. The net force will also lie in this direction at this instant. A force triangle and trigonometry can be used here.



$$\Sigma F = 1.47 \tan 60.0^\circ = 2.55 \text{ N towards the left}$$

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

$$F_t = \frac{1.47}{\cos 60.0^\circ} = 2.94 \text{ N along direction of the cord}$$

- e** $\Sigma F = 2.55 \text{ N}$, $m = 0.150 \text{ kg}$, $r = 1.30 \text{ m}$, $v = ?$

$$\Sigma F = \frac{mv^2}{r}$$

$$2.55 = \frac{0.150v^2}{1.30}$$

$$\Rightarrow v = 4.70 \text{ m s}^{-1}$$

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.

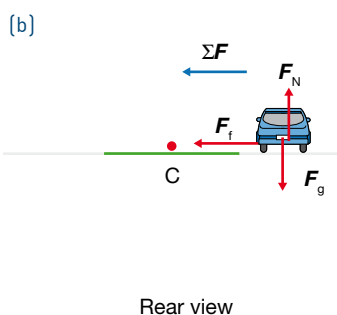
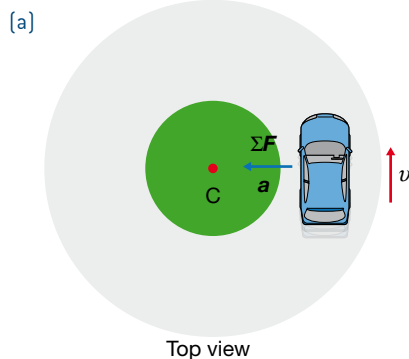


Figure 2.39 (a) The car is travelling in a circular path on a horizontal track. The acceleration and net force are towards C. (b) The vertical forces balance and it is friction between the tyres and the road that enables the car to corner.



Figure 2.38 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.

Consider a car travelling anticlockwise around a horizontal roundabout with a constant speed v . As can be seen in Figure 2.39a, the car has an acceleration towards C and so the net force is also towards C. In Figure 2.39b, the forces acting on the car can be seen. The vertical forces (i.e. gravity and the normal force) are balanced. The only horizontal force is the sideways force that the road exerts on the car tyres. This is a force of friction, F_f , and is unbalanced, so this is equal to the net force, ΣF . 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 in a straight line 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.

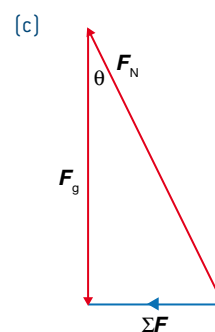
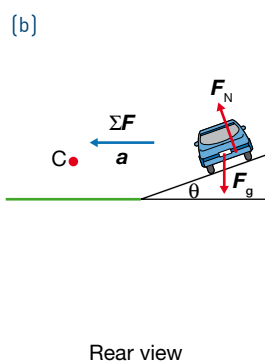
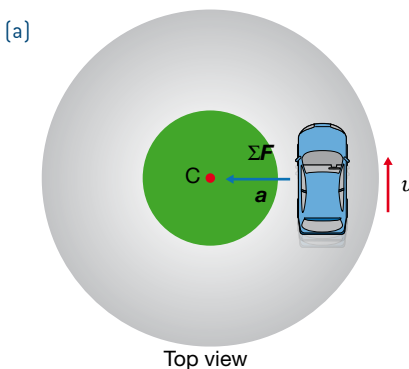


Figure 2.40 (a) The car is travelling in a circular path on a banked track. The acceleration and net force are towards C. (b) The banked track means that the normal force has an inwards component. This is what enables the car to turn the corner. (c) Vector addition gives the net force as horizontally towards the centre.

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 weight, F_g , and the normal force, F_N , from the track. As can be seen in Figure 2.40b, these forces are unbalanced.

They add together to give a net force that is horizontal and directed towards C. If the angle and weight are known, trigonometry can be used to calculate the net force (Figure 2.40c) and so determine the design speed. It is worth noting that the normal force will be larger here than on a flat track. The rider and bike would feel a larger force acting from the road.

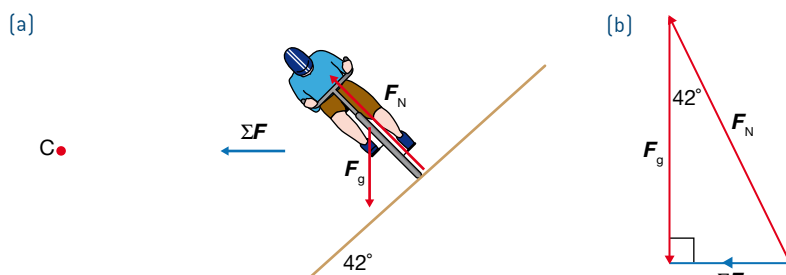
Worked example 2.6B

A curved section of track on an Olympic velodrome has radius of 50 m and is banked at an angle of 42° to the horizontal.

- Calculate the net force acting on a cyclist riding at the design speed.
- At what speed would a cyclist of mass 75 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

- The forces acting on the cyclist are gravity and the normal force from the track. The net force is horizontal and 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.

$$\begin{aligned}\tan 42^\circ &= \frac{\Sigma F}{F_g} \\ 0.90 &= \frac{\Sigma F}{75 \times 9.8} \\ \Rightarrow \Sigma F &= 660 \text{ N towards C}\end{aligned}$$

- The net force is a centripetal force, so:

$$\begin{aligned}\Sigma F &= \frac{mv^2}{r} \\ 660 &= \frac{75 \times v^2}{50} \\ \Rightarrow v &= \sqrt{441} = 21 \text{ m s}^{-1}\end{aligned}$$

The cyclist in Worked example 2.6B, travelling on this section of track at 21 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.

However, what would happen if they slowed down? Gravity would not change, but now both the normal force and net 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.

Physics file



Figure 2.41 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.



INTERACTIVE TUTORIAL

High-speed cornering

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 2.42b) are unbalanced. The forces are the weight force, F_g , and the force from the track. The track exerts a reaction force, F_r , on the rider that acts both inwards and upwards. The inwards component is the frictional force, F_f , between the track and the tyres. The upwards component is the normal force, F_N , 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 net force is directed towards C. By analysing the vertical and horizontal components in Figure 2.42b, we see that the weight force, F_g , must balance the normal force, F_N . The net force that is producing the centripetal acceleration is supplied by the frictional force, 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.

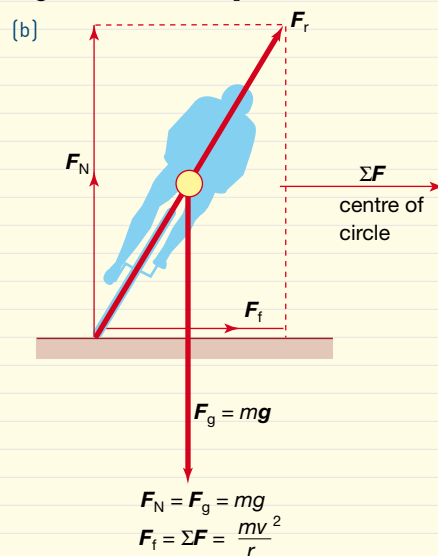


Figure 2.42 (a) Australia's Casey Stoner won the 2007 MotoGP championship. Here he is leaning his bike as he takes a corner at Phillip Island. Leaning into the corner enables him to corner at higher speeds. In fact, the bike would flip if he did not lean it. (b) The forces acting as the rider turns a corner are the weight, F_g , the normal force, F_N , and the friction, F_f , between the tyres and the road. The friction supplies the unbalanced force that leads to the circular motion.



2.6 summary

Aspects of circular motion

- 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 net force is the sum of the tension and weight forces. 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 from the track. They add to give a net force towards the centre of the circular motion.

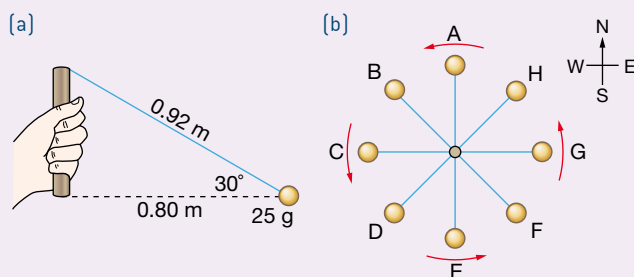


2.6 questions

Aspects of circular motion

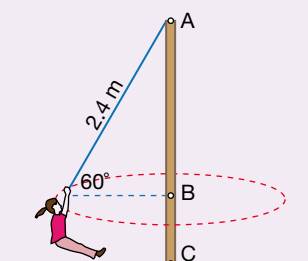
The following information applies to questions 1–7.

During a high-school physics experiment, a copper ball of mass 25.0 g was attached to a very light piece of steel wire 0.920 m long and whirled in a circle at 30.0° to the horizontal. The ball moves in a circular path of radius 0.800 m with a period of 1.36 s. The top view of the resulting motion of the ball is shown in (b).



- What is the direction of the ball's velocity at point:
 - A?
 - B?
 - G?
- What is the direction of the ball's acceleration when it is at point:
 - A?
 - B?
 - G?
- What is the direction of the net force that acts on the ball when it is at point:
 - A?
 - D?
 - F?
- If the wire had snapped when the ball was at point E, what would be the direction of the ball's subsequent velocity?
- Which one of the following statements is correct?
 - The ball's kinetic energy and momentum are both changing.
 - The ball's kinetic energy and momentum are both constant.
 - The ball's kinetic energy is constant but its momentum is changing.
 - The ball's momentum is constant but its kinetic energy is changing.

- Calculate the orbital speed of the ball.
 - What is the centripetal acceleration of the ball?
 - What is the magnitude of the centripetal force acting on the ball?
- Draw a diagram similar to diagram (a) that shows all the forces acting on the ball at this time.
 - What is the magnitude of the tension in the wire?
- A child of mass 30 kg is playing on a maypole swing in a playground. The rope is 2.4 m long and at an angle of 60° to the horizontal as she swings freely in a circular path. Ignore the mass of the rope in your calculations.



- Calculate the radius of her circular path.
 - Identify the forces that are acting on her as she swings freely.
 - What is the direction of her acceleration when she is at the position shown in the diagram?
 - Calculate the net force acting on the girl.
 - 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 1200 kg is being driven around the track at 18 m s^{-1} .
- Calculate the magnitude of the net force acting on the car (in kN).
 - Calculate the angle to the horizontal at which the track is banked.
 - The driver now drives around the track at 30 m s^{-1} . What would the driver have to do maintain their circular path around the track?

2.7

Circular motion in a vertical plane

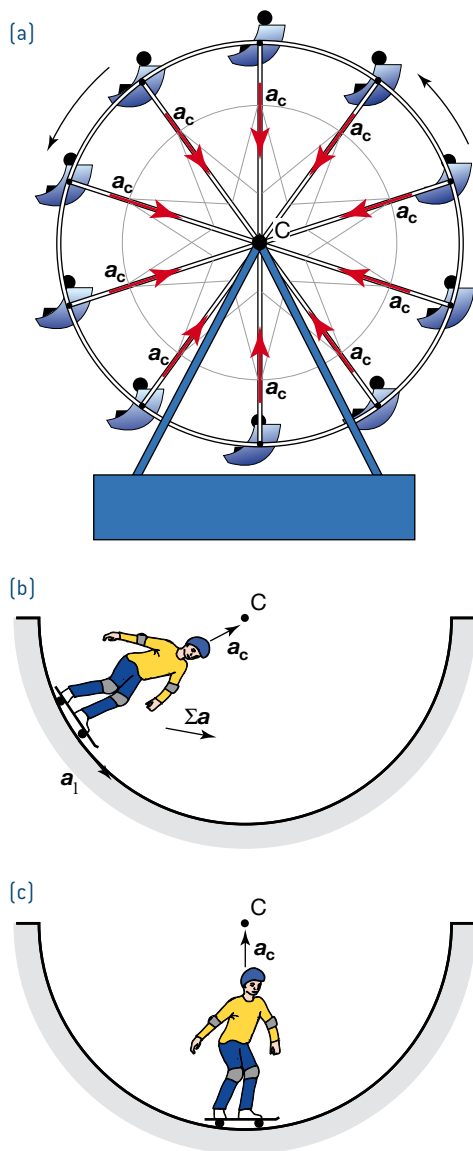


Figure 2.43 (a) The seats in a ferris wheel move with a constant speed and so have a constant centripetal acceleration directed towards the centre C. (b) The skateboarder is speeding up and so has both a linear and a centripetal acceleration. The net acceleration, $\Sigma \mathbf{a}$, is not towards 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.

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 2.5 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 2.43a).

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 gravitational potential energy is converted into kinetic energy. This means there will be a linear acceleration, \mathbf{a}_l , as well as a centripetal acceleration, \mathbf{a}_c . The resultant acceleration cannot be directed towards the centre of the circular path (Figure 2.43b). 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 2.43c). The same applies at the very top of a circular path. For this reason, motion at these points is easier to analyse.

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 4.0 m s^{-1} . If the person's mass is 50 kg and the gravitational field strength is 9.8 m s^{-2} , the forces acting on the person can be calculated. These forces are the weight, \mathbf{F}_g or \mathbf{W} , and the normal force, \mathbf{F}_N or \mathbf{N} , from the seat (Figure 2.44). The person is moving in a straight line with a constant speed and so there is no unbalanced force acting. The weight force balances the normal reaction force from the seat. The normal force is therefore 490 N up.

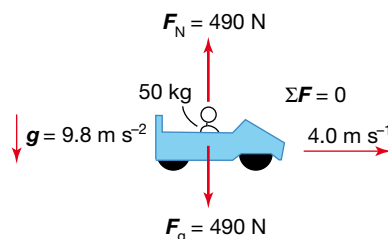


Figure 2.44 The vertical forces are in balance in this situation, i.e. $F_N = F_g$.

Now consider the forces that act on the person after the cart has reached the bottom of a circular dip of radius 2.5 m and is now moving at 8.0 m s^{-1} (Figure 2.45). The person will have a centripetal acceleration due to the circular path. This centripetal acceleration is directed towards the centre C of the circular path—in this case, vertically upwards. The person's centripetal acceleration is:

$$\mathbf{a} = \frac{v^2}{r} = \frac{8.0^2}{2.5} = 26 \text{ m s}^{-2} \text{ towards C, i.e. upwards}$$

and the net (centripetal) force acting on the person is given by:

$$\Sigma \mathbf{F} = m\mathbf{a} = 50 \times 26 = 1300 \text{ N upwards}$$

The normal force, F_N , and the weight force, F_g , are now *not* in balance. They add together to give an upwards force of 1300 N. This indicates that the normal force must be greater than the weight force by 1300 N. In other words, the normal force is $490 \text{ N} + 1300 \text{ N} = 1790 \text{ N}$ up.

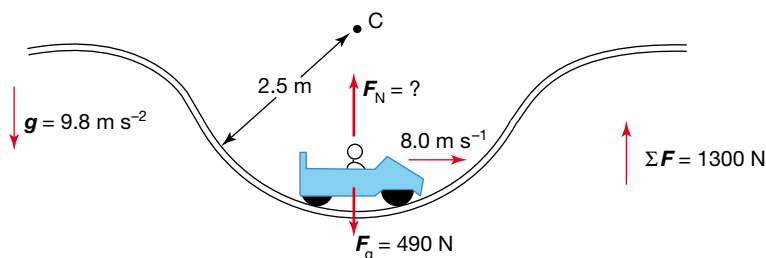


Figure 2.45 The person has a centripetal acceleration that is directed upwards, and so the net force is also upwards. So here, the magnitude of the normal force, F_N , is greater than the weight, F_g , a situation that is felt by the rider.

Now consider the situation as the cart moves over the top of a hump of radius 2.5 m with a lower speed of 2.0 m s^{-1} (Figure 2.46). The person now has a centripetal acceleration that is directed vertically downwards towards the centre C of the circle. Therefore, the net force acting at this point is also directed vertically downwards. The centripetal acceleration is:

$$a = \frac{v^2}{r} = \frac{2.0^2}{2.5} = 1.6 \text{ m s}^{-2} \text{ downwards}$$

and the net (centripetal) force is therefore:

$$\Sigma F = ma = 50 \times 1.6 = 80 \text{ N downwards}$$

The weight force and the normal force are again not in balance. They add to give a net force of 80 N down. The weight force, F_g , must therefore be 80 N greater than the normal force, F_N . This tells us that the normal force is:

$$490 \text{ N} + -80 \text{ N} = 410 \text{ N up}$$

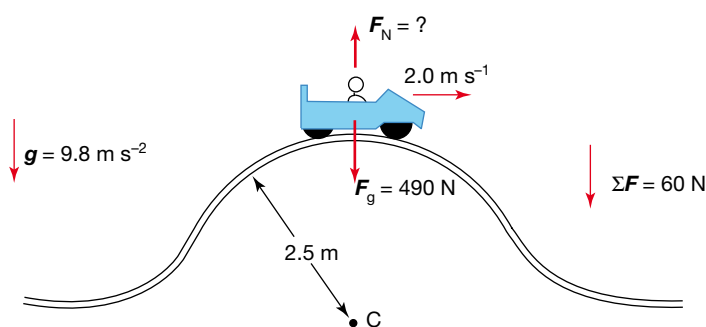


Figure 2.46 The centripetal acceleration is downwards, and so the net force is also in that direction. At this point, the magnitude of the normal force, F_N , is less than the weight, F_g , of the person.

It is interesting to compare the normal forces that act on the person in these situations. The normal force when travelling horizontally is 490 N upwards. At the bottom of the dip, the normal force is 1790 N upwards. In other words, the seat pushes into the person with a greater force than usual in the dip. This makes the person feel heavier than normal. If the person had been sitting on weighing scales at this time, it would have shown a higher than usual reading. Conversely, the person feels a much smaller force, 410 N, from the seat when travelling over the top of the hump. This gives the sensation of feeling lighter than usual.



PRACTICAL ACTIVITY 14

Investigating circular motion in a vertical plane

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 normal force acting on them varies throughout the ride. The normal force acting on the person gives their *apparent weight*. This will also 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 2.46, but at an increased speed: 4.95 m s^{-1} . This speed would give a centripetal acceleration of:

$$a = \frac{v^2}{r} = \frac{4.95^2}{2.5} = 9.8 \text{ m s}^{-2} \text{ towards C}$$

The centripetal (i.e. net) force acting in this case would be:

$$\Sigma F = ma = 50 \times 9.8 = 490 \text{ N downwards}$$



Figure 2.47 When a car travels too fast for the radius of curvature of the road, the normal force, F_N , from the road is reduced to zero and the car can lift off the road surface.

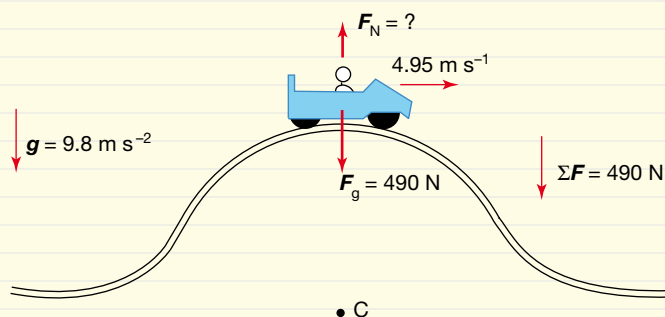


Figure 2.48 At this speed the centripetal force, ΣF , is equal to the weight, F_g , of the person. In other words, the normal force, F_N , is zero. The person would feel as though they were at the point of lifting off the seat.

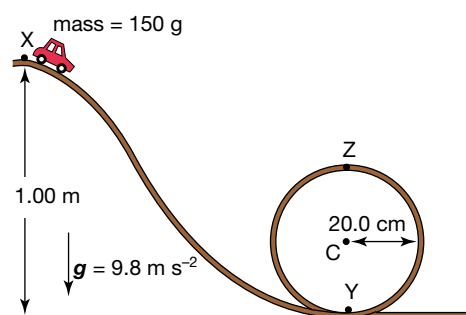
By considering the force vectors, the weight fully accounts for the centripetal force. In other words, the normal force, F_N , is now zero. While travelling over the hump, there would be no interaction between the person and the seat, and they would be at the point of lifting off. If the speed of the cart had been any greater, the person would have left the seat and might have needed a restraint to remain in the cart. 'Lift-off' occurs when the only force acting on the body is gravity, i.e. when it moves with an acceleration of 9.8 m s^{-2} down.

Worked example 2.7A

A student arranges a toy car track with a vertical loop of radius 0.200 m, as shown.

A car of mass 150 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 g is 9.80 m s^{-2} , and ignoring friction, calculate:

- the speed of the car as it reaches the bottom of the loop, point Y
- the centripetal acceleration of the car at point Y
- the normal reaction force from the track at point Y
- the speed of the car as it reaches the top of the loop at point Z
- the apparent weight of the car at point Z
- the 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:

$$\begin{aligned}\Sigma E &= E_k + U_g = \frac{1}{2}mv^2 + mgh \\ &= 0 + 0.150 \times 9.80 \times 1.00 \\ &= 1.47 \text{ J}\end{aligned}$$

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.

$$\text{i.e. } \frac{1}{2}mv^2 = 1.47$$

$$\therefore 0.5 \times 0.150 \times v^2 = 1.47$$

$$\therefore v = \sqrt{19.6} = 4.43 \text{ m s}^{-1}$$

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

$$\begin{aligned}a &= \frac{v^2}{r} \\ &= \frac{4.43^2}{0.200} \\ &= 98.1 \text{ m s}^{-2} \text{ towards C, i.e. upwards}\end{aligned}$$

- c** The centripetal force on the car at Y is: $\Sigma F = ma = 0.150 \times 98.1 = 14.7 \text{ N}$ upwards, towards C. The normal force, F_N , can be determined by vector subtraction:

$$\begin{aligned}\Sigma F &= F_N + F_g \\ \therefore 14.7 \text{ up} &= F_N + 1.47 \text{ down} \\ \therefore F_N &= 14.7 \text{ up} - 1.47 \text{ down} = 16.2 \text{ N upwards}\end{aligned}$$

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 car has a gravitational potential energy of:

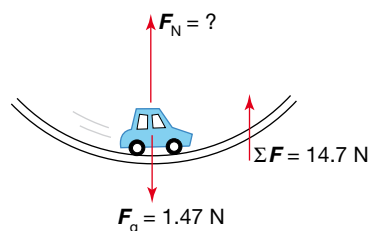
$$U_g = mgh = 0.150 \times 9.80 \times 0.400 = 0.588 \text{ J}$$

$$\begin{aligned}\Sigma E &= E_k + U_g \\ \therefore 1.47 &= \frac{1}{2}mv^2 + 0.588 \\ \therefore 0.88 &= 0.5 \times 0.150 \times v^2 \\ \therefore v &= \sqrt{11.7} = 3.42 \text{ m s}^{-1}\end{aligned}$$

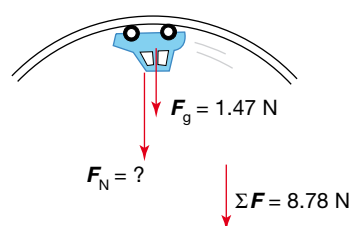
- e** The centripetal acceleration is:

$$\begin{aligned}a &= \frac{v^2}{r} \\ &= \frac{3.42^2}{0.200} \\ &= 58.5 \text{ m s}^{-2} \text{ downwards, towards C}\end{aligned}$$

at point Y



at point Z



Physics file

A fighter pilot in a loop manoeuvre can safely experience centripetal accelerations of up to around $5g$, 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 net force at point Y is:

$$\begin{aligned}\Sigma F &= ma \\ &= 0.150 \times 58.5 \\ &= 8.78 \text{ N downwards}\end{aligned}$$

The forces acting at point Z are shown in diagram (b). The apparent weight is the normal force, F_N , and can be found by:

$$\begin{aligned}\Sigma F &= F_N + F_g \\ \therefore 8.78 \text{ down} &= F_N + 1.47 \text{ down} \\ \therefore F_N &= 7.31 \text{ N down. This is about five times greater than the actual weight.}\end{aligned}$$

- f** When the car is travelling at the speed at which it just loses contact with the track, the normal force, F_N , i.e. the apparent weight, is zero. This means that the centripetal force on the car equals its weight.

$$\begin{aligned}\Sigma F &= F_g \\ \frac{mv^2}{r} &= mg \\ v &= \sqrt{gr} \\ &= \sqrt{9.80 \times 0.200} = 1.40 \text{ m s}^{-1}\end{aligned}$$

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

$$\begin{aligned}\Sigma E &= E_k + U_g \\ &= (0.5 \times 0.150 \times 1.40^2) + (0.150 \times 9.80 \times 0.400) \\ &= 0.147 + 0.588 \\ &= 0.735 \text{ J}\end{aligned}$$

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

$$\begin{aligned}\therefore mgh &= 0.735 \\ \therefore 0.150 \times 9.80 \times h &= 0.735, \text{ so } h = 0.500 \text{ m}\end{aligned}$$

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 2.49, 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 2.49 The thrill seekers on this roller-coaster ride don't fall out when upside down because the centripetal acceleration of the cart is greater than 9.8 m s^{-2} down.

The reason people don't fall out is that their centripetal acceleration while on the roller-coaster is greater than the acceleration due to gravity (9.8 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 a downwards acceleration in excess of 9.8 m s^{-2} , and continually exerting a normal force on the eraser. This acceleration of 9.8 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 don't fall out at the top because the motion of the roller-coaster gives them a centripetal acceleration that is greater than 9.8 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 2.50). 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 9.8 m s^{-2} . At the critical speed, the normal force, F_N , on the person will be zero. In other words, the seat will exert no force on the person at this speed. The centripetal force, ΣF , is:

$$\Sigma F = F_g + F_N \text{ but } F_N = 0, \text{ so}$$

$$\Sigma F = F_g$$

$$\therefore \frac{mv^2}{r} = mg$$

$$\therefore v = \sqrt{gr} = \sqrt{9.8 \times 15} = 12 \text{ m s}^{-1}$$

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.

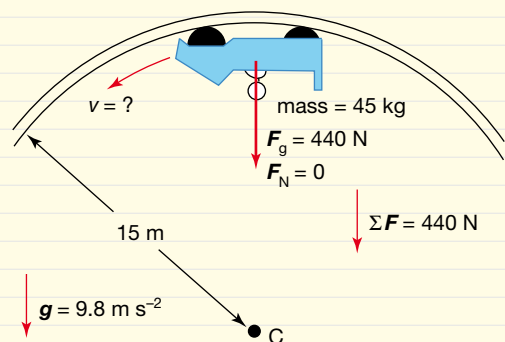


Figure 2.50 At the speed at which the normal force, F_N , becomes zero, the net force, ΣF , on the rider will equal the weight, F_g , and they will move so that the centripetal acceleration is 9.8 m s^{-2} down.



2.7 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 force that is acting. This may be different from the actual weight.



2.7 questions

Circular motion in a vertical plane

In the following questions, assume that $g = 9.80 \text{ m s}^{-2}$ and ignore the effects of air resistance.

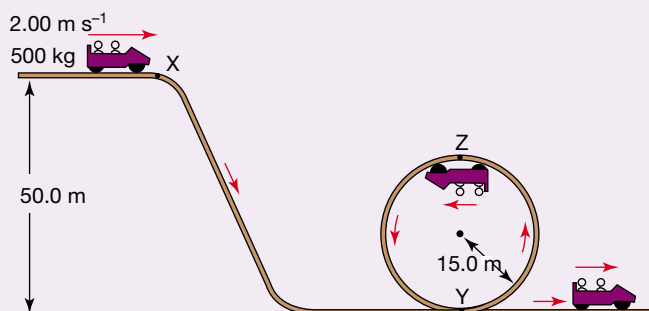
The following information applies to questions 1 and 2.

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 g and the radius of the circle is 1.5 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 kg encounters a speed hump of radius 10 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 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?

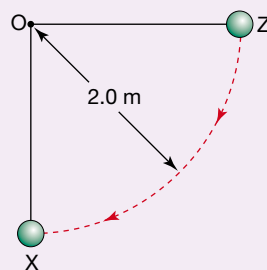
The following information applies to questions 6 and 7.

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 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 m .

The following information applies to questions 9 and 10.

A light wire of length 2.0 m supports a bowling ball of mass 4.0 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 review

For the following questions, assume that $g = 9.8 \text{ m s}^{-2}$ and ignore the effects of air resistance.

- 1 A ball of mass 200 g is dropped from a vertical height of 10 m onto a horizontal concrete floor. The ball rebounds upwards with an initial vertical velocity of 10 m s^{-1} . The time of interaction for this impact is 1.0 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 1.0 kg m s^{-1} up
C 2.0 kg m s^{-1} up	D 10 kg m s^{-1} up
 - c What is the size of the impulse that has acted on the ball as it bounces?

A 2.8 N s	B 2.0 N s
C 0.8 N s	D 4.8 N s
 - d Calculate the average net force that has acted on the ball.

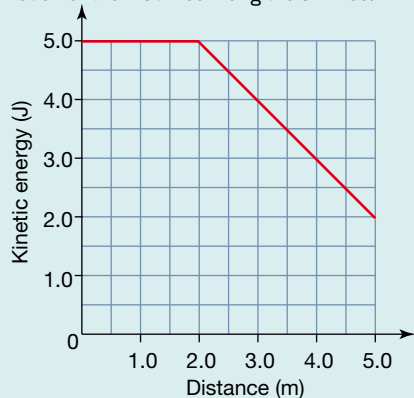
The following information applies to questions 2 and 3.

A student supplies a constant force of 200 N at an angle of 60° 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} .

- 2
 - 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?
- 3
 - 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 4–6.

An ice-puck of mass of 200 g moves across a horizontal surface 5.0 m long. One section of the surface is frictionless while the other is rough. The following graph shows the kinetic energy of the puck as a function of the distance along the surface.



- 4 Which section of the surface is rough? Justify your answer.
- 5 What is the magnitude of the frictional force that acts on the puck when it is on the rough surface?
- 6 How much heat energy is produced by the frictional force?

A 3.0 J
B 5.0 J
C 2.0 J
D 7.0 J

The following information applies to questions 7–10.

A 50 kg boy stands on a 200 kg sled that is at rest on a frozen pond. The boy jumps off the sled with a velocity of 4.0 m s^{-1} east.

- 7
 - 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 4.4 m s^{-1} west.

- 8 Assuming the pond surface is frictionless, what is the velocity of the sled just before he jumps on?
- 9 What is the speed of the boy once he is on the sled?
- 10 As the boy jumps on the sled, what change in momentum is experienced by:
 - a the sled?
 - b the boy?
- 11 A 250 g snooker ball travelling at 10 m s^{-1} collides with a stationary 100 g snooker ball. Assuming that this collision has an energy efficiency of 95%, what is the total kinetic energy of the balls after the collision?

A 95 J
B 13 J
C 6.0 J
D 12 J

The following information applies to questions 12–14.

Two air-track gliders, both travelling at 2.0 m s^{-1} , approach each other on an air-track. The gliders, with masses of 300 g and 100 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.

- 12 Calculate their common velocity after colliding.
- 13 What is the energy efficiency of this collision?

A 100%
B 0%
C 25%
D 75%

14 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 15 and 16.

A homemade tennis-ball server consists of a spring-loaded plunger (of negligible mass) inside a length of cylindrical tube. The physics student who designed this device used it to launch a tennis ball of mass 50 g. The spring was compressed by 10 cm and then released, resulting in the ball leaving the tubing with initial horizontal velocity of 8.0 m s^{-1} .

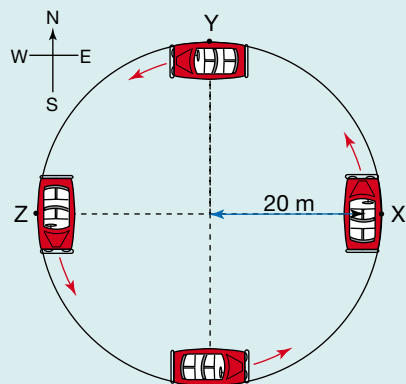
15 If the spring constant of the plunger is 2000 N m^{-1} , calculate the energy that was transformed into heat as a result of friction inside the tubing.

16 For a spring compression of 10 cm, which one or more of the following modifications would result in the ball having a greater velocity on leaving the tube?

- A Reduce the length of the tube.
- B Increase the length of the tube.
- C Increase the spring constant for the spring.
- D Use a tennis ball with a lower mass.

The following information applies to questions 17–20.

A car of mass 1500 kg is driven at constant speed of 10 m s^{-1} around a level, circular roundabout. The centre of mass of the car is always 20 m from the centre of the track.



17 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:

- a point X?
- b point Y?
- c point Z?

18 What is the period of revolution for this car?

- 19
- 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 Calculate the unbalanced frictional force acting on the tyres at point Z.

20 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.

The following information applies to questions 21 and 22.

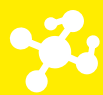
A cycling velodrome has a turn that is banked at 33° to the horizontal. The radius of the track at this point is 28 m.

21 Determine the speed (in km h^{-1}) at which a cyclist of mass 55 kg would experience no sideways force on their bike as they ride this section of track.

22 Calculate the size of the normal force that is acting on the cyclist. How does this compare with the normal force if they were riding on a flat track?

23 The ferris wheel at an amusement park has an arm radius of 10 m and its compartments move with a constant speed of 5.0 m s^{-1} .

- a Calculate the normal force that a 50 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.



Gravity and satellites



Through the ages, people have looked to the skies and wondered about the motion of the celestial bodies: how the path of the Sun seems to change from season to season; how the appearance of the Moon changes during each 'moonth'; how the stars move nightly across the sky; and the more complicated motion of the planets. The efforts and insights of people such as Aristotle, Ptolemy, Copernicus, Kepler, Galileo and Newton have, over many centuries, led to our present understanding of the motion of the stars, especially our Sun, its planets and their moons and our place in the Universe. The Universe consists of perhaps 100 billion galaxies that are vast distances apart in space. The Universe is expanding—the galaxies are moving further apart from each other and this expansion is in fact accelerating. The reason for this acceleration is currently a topic of major dispute among cosmologists. Humans have barely begun to explore the Universe. The Space Age began in 1957 when the then Soviet Union launched the first artificial satellite, Sputnik I. Since then, we have made remarkable progress in space technology. The International Space Station (ISS) is currently being assembled by astronauts working 400 km above the Earth's surface. The first module was placed in orbit in 1998 and the first astronauts moved in 2 years later. When the ISS is completed in 2010, it will be almost 100 m long.

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 empirical approach used earlier by Johannes Kepler, who attempted to fit rules to match the data without trying to provide an explanation for these rules.



by the end of this chapter

you will have covered material from the study of gravity and satellites, including:

- Newton's law of universal gravitation
- gravitational fields
- satellite motion
- energy transfers from $F-d$ and $g-d$ graphs
- apparent weight, weightlessness and apparent weightlessness.



CHAPTER 3

3.1

Newton's law of universal gravitation

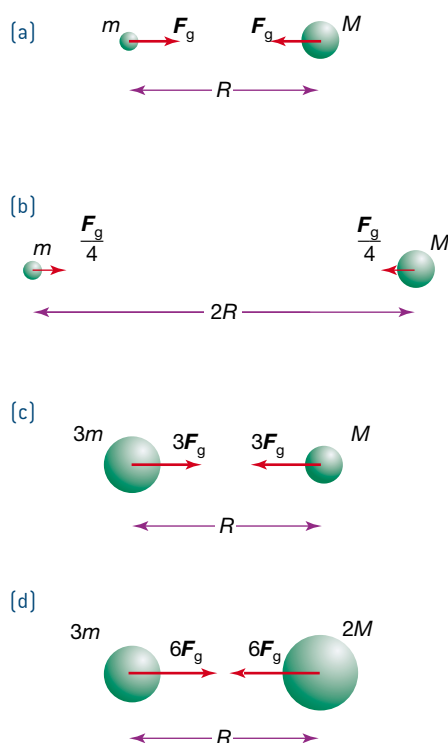


Figure 3.1 The gravitational force of attraction between two bodies depends on their separation and each of their masses.

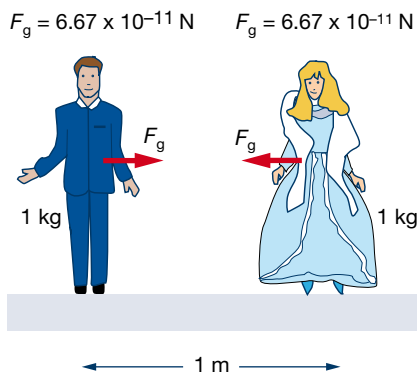


Figure 3.2 The gravitational force of attraction between these two dolls is so weak that it does not cause them to be thrown together.

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.

From the laws of motion, 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.

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

- is one of attraction, and acts from the centre of each mass
- acts equally 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. $\mathbf{F} \propto \frac{1}{R^2}$, where R is the distance 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.



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

$$F = \frac{GMm}{R^2}$$

where F = the gravitational force acting on each body (N)

M and m are the masses of the bodies (kg)

R = the distance between the centres of the bodies (m)

The universal gravitational constant G is equal to $6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$.

The extremely small value of the universal gravitational constant gives an indication of why gravitational forces are not noticeable between everyday objects. Consider two 1.0 kg dolls whose centres are 1 m apart (Figure 3.2). The gravitational force of attraction, \mathbf{F} , between these bodies is given by:

$$\begin{aligned} F &= \frac{GMm}{R^2} \\ &= \frac{6.67 \times 10^{-11} \times 1.0 \times 1.0}{1.0^2} \\ &= 6.67 \times 10^{-11} \text{ N} \end{aligned}$$

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 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 6400 km or 6.4×10^6 m. Given that the mass of the Earth, M_E , is 6.0×10^{24} kg, the gravitational force of attraction that the 1 kg mass and the Earth exert on each other is:

$$\begin{aligned} F &= \frac{GM_E m}{R^2} \\ &= \frac{6.67 \times 10^{-11} \times 6.0 \times 10^{24} \times 1.0}{(6.4 \times 10^6)^2} \\ &= 9.8 \text{ N} \end{aligned}$$

This value is the *weight* of the 1 kg mass. This weight force of 9.8 N acts on *both* the 1 kg mass and the Earth (Newton's third law). However, the effect of the force is vastly different. If the 1 kg mass is free to fall, it will accelerate at 9.8 m s^{-2} towards the Earth, whereas the Earth with its enormous mass, will not be moved to any measurable degree by this 9.8 N force.

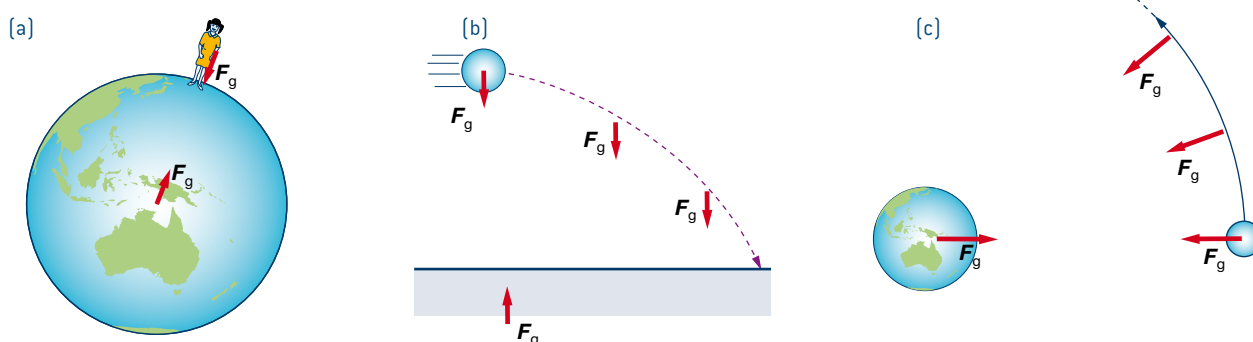


Figure 3.3 The effect of the gravitational force is significant when one of the masses involved is large. Gravitation is the force that (a) pulls you towards the Earth, (b) causes a projectile to fall towards the Earth, and (c) causes the Moon to 'fall' around the Earth.

Gravitation, although it is the weakest of the four fundamental forces that exist in nature (see Physics in action, page 82), is the most far-reaching of these forces; its influence stretches across the Universe.

Worked example 3.1A

A 10.0 kg watermelon falls a short distance to the ground. If the Earth has a radius of 6.4×10^6 m (6400 km) and a mass of 6.0×10^{24} kg, calculate:

- the gravitational force that the Earth exerts on the watermelon
- the gravitational force that the watermelon exerts on the Earth
- the acceleration of the watermelon towards the Earth
- the acceleration of the Earth towards the watermelon.

Solution

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

$$\begin{aligned} F &= \frac{GMm}{R^2} \\ &= \frac{6.67 \times 10^{-11} \times 6.0 \times 10^{24} \times 10.0}{(6.4 \times 10^6)^2} \\ &= 98 \text{ N} \end{aligned}$$

- b** The watermelon also exerts a force of attraction of 98 N 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.

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.

- c** The acceleration of the watermelon is:

$$a = \frac{\sum F}{m_{\text{melon}}}$$

$$= \frac{98}{10}$$

$$= 9.8 \text{ m s}^{-2} \text{ towards the centre of the Earth}$$

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

- d** The acceleration of the Earth is:

$$a = \frac{\sum F}{m_{\text{Earth}}}$$

$$= \frac{98}{6.0 \times 10^{24}}$$

$$= 1.6 \times 10^{-23} \text{ m s}^{-2} \text{ towards the centre of the watermelon}$$

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

Worked example 3.1B

The gravitational force that acts on a 1200 kg space probe on the surface of Mars is $4.43 \times 10^3 \text{ N}$. The radius of Mars is 3400 km. Without using the mass of Mars, determine the gravitational force that acts on the space probe when it is:

- a** 3400 km above the surface of Mars
b 6800 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:

$$\text{i.e. } F \propto \frac{1}{R^2}$$

When it is sitting on the surface of Mars, the probe is 3400 km from the centre of the planet and the gravitational force acting on it is $4.43 \times 10^3 \text{ N}$. When the probe is at an altitude of 3400 km, it is 6800 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:

$$\begin{aligned} \text{i.e. } F &= \frac{4.43 \times 10^3}{2^2} \\ &= \frac{4.43 \times 10^3}{4} \\ &= 1.11 \times 10^3 \text{ N} \end{aligned}$$

The inverse square nature of the relationship between force and distance can also be used to determine the force:

$$\begin{aligned} \frac{F_2}{F_1} &= \frac{R_1^2}{R_2^2} \\ \therefore \frac{F_2}{4.43 \times 10^3} &= \frac{3400^2}{6800^2} \end{aligned}$$

$$\therefore F_2 = 1.11 \times 10^3 \text{ N}$$

If this ratio approach is used, it is not necessary to use standard SI units, but the units used for each of the quantities must be the same.

- b** At 6800 km above the surface of Mars, the probe is 10 200 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:

$$\begin{aligned} \text{i.e. } F &= \frac{4.43 \times 10^3}{3^2} \\ &= \frac{4.43 \times 10^3}{9} \\ &= 492 \text{ N} \end{aligned}$$

The ratio of the forces and distances can again be used to find the size of this force:

$$\frac{F_2}{F_1} = \frac{R_1^2}{R_2^2}$$

$$\therefore \frac{F_2}{4.43 \times 10^3} = \frac{3400^2}{10\,200^2}$$

$$\therefore F_2 = 492 \text{ N}$$

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 3.4 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 $6.4 \times 10^6 \text{ m}$, and the gravitational force on a 1 kg mass were used to calculate the Earth's mass:

$$F = \frac{GM_E M}{R^2}$$

$$\therefore 9.8 = \frac{6.67 \times 10^{-11} \times M_E \times 1.0}{(6.4 \times 10^6)^2}$$

$$\therefore M_E = \frac{9.8 \times (6.4 \times 10^6)^2}{6.67 \times 10^{-11} \times 1.0}$$

$$= 6.0 \times 10^{24} \text{ kg}$$

The value for G that was calculated later was only 1% different from the value that we use today.

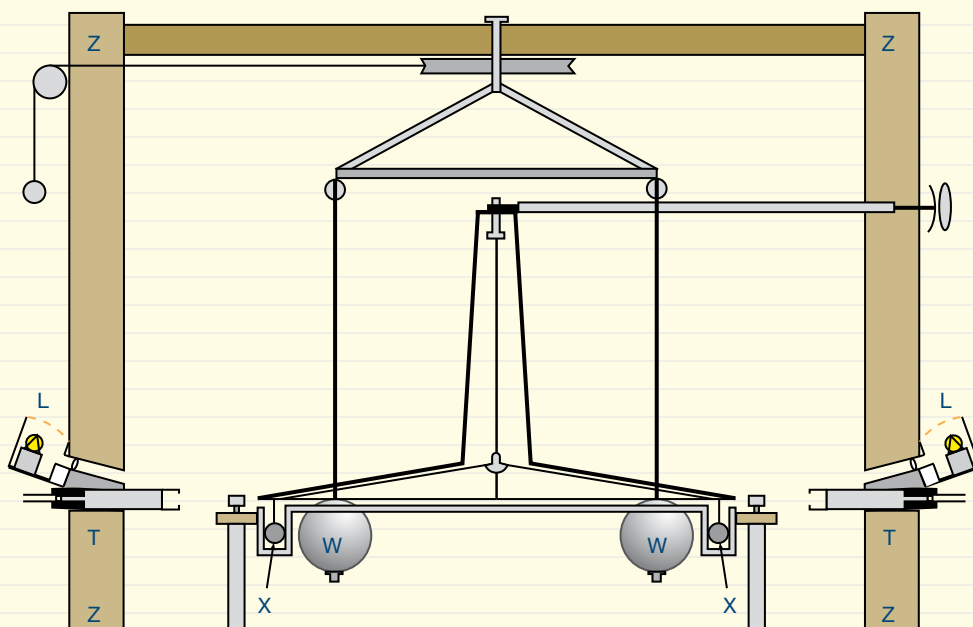


Figure 3.4 A diagram of the apparatus that was used by English physicist Henry Cavendish in 1798 to experimentally determine the value of the Earth's density.

Gravitation: 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,



Figure 3.5 Gravitational forces hold planets in their orbits around the Sun.

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.



3.1 summary

Newton's law of universal gravitation

- Gravitation is a force of attraction that acts between all bodies.
- The gravitational force acts equally on each of the bodies.
- The gravitational force is directly related to the masses:

$$F \propto M, m$$

- 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:

$$F \propto \frac{1}{R^2}$$

- The separation distance of the bodies is measured from the centre of mass of each object.
- Newton's law of universal gravitation is:

$$F = \frac{GMm}{R^2}$$

- The constant of universal gravitation is:
 $G = 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$



3.1 questions

Newton's law of universal gravitation

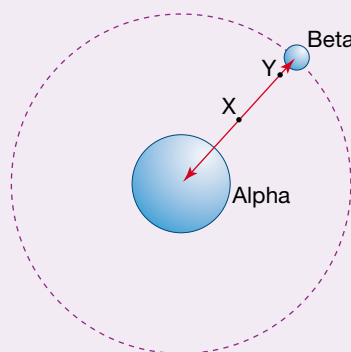
For these questions, assume that $G = 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$, and that distances are measured between the centres of the bodies.

- 1 For this question, assume that:
 mass of Earth = $6.0 \times 10^{24} \text{ kg}$
 radius of Earth = $6.4 \times 10^6 \text{ m}$
 mass of the Moon = $7.3 \times 10^{22} \text{ kg}$
 mean radius of Moon's orbit = $3.8 \times 10^8 \text{ m}$
 Calculate the gravitational force of attraction that exists between the following objects.
 - a A 100 g apple and a 200 g orange that are 50 cm apart
 - b A pair of 1000 kg boulders that are 10.0 m apart
 - c The Earth and a satellite of mass 20 000 kg in orbit at an altitude of 600 km
 - d The Moon and the Earth
 - e The Earth and a 60 kg student standing on its surface
 - f A proton of mass $1.67 \times 10^{-27} \text{ kg}$ and an electron of mass $9.11 \times 10^{-31} \text{ kg}$ separated by $5.30 \times 10^{-11} \text{ m}$ in a hydrogen atom
- 2 Newton's law of universal gravitation defines the symbol G as a 'universal constant'.
 - a Explain the meaning of this term and its significance.
 - b Identify another universal constant.
- 3 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.
- 4 Mars has two natural satellites. Phobos has a mean orbital radius of $9.4 \times 10^6 \text{ m}$, while the other moon, Deimos, is located at a mean distance of $2.35 \times 10^7 \text{ m}$ from the centre of Mars. Data: mass of Mars = $6.42 \times 10^{23} \text{ kg}$; mass of Phobos = $1.08 \times 10^{16} \text{ kg}$; mass of Deimos = $1.8 \times 10^{15} \text{ kg}$.
 Calculate the value of the ratio:

$$\frac{\text{gravitational force exerted by Mars on Phobos}}{\text{gravitational force exerted by Mars on Deimos}}$$

- 5 A comet of mass 1000 kg is plummeting towards Jupiter. Jupiter has a mass of $1.90 \times 10^{27} \text{ kg}$ and a planetary radius of $7.15 \times 10^7 \text{ m}$. If the comet is about to crash into Jupiter, calculate the following:
 - a the magnitude of the gravitational force that Jupiter exerts on the comet
 - b the magnitude of the gravitational force that the comet exerts on Jupiter
 - c the acceleration of the comet towards Jupiter
 - d the acceleration of Jupiter towards the comet.
- 6 An astronaut travels away from Earth to a region in space where the gravitational force due to Earth is only 1.0% 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 .



- 7 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?
- 8 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 ?

3.2

Gravitational fields

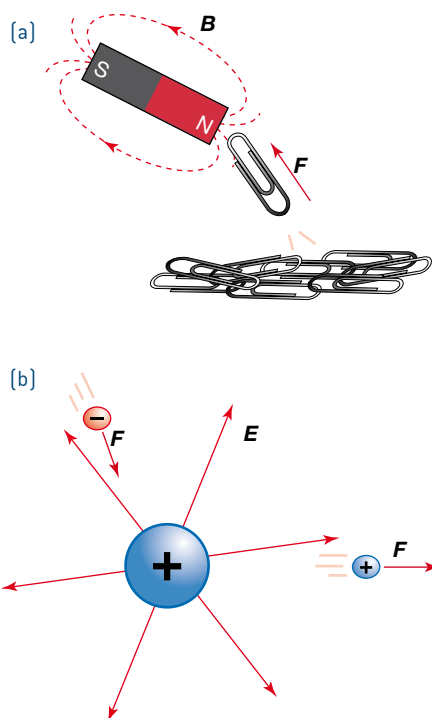


Figure 3.6 A field is a region where a body will experience force. (a) The paper clip experiences a force because it is in the magnetic field B of the bar magnet. (b) The small charged particles experience forces because they are in the electric field E of the central charge. These fields are invisible.

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. As illustrated in Figure 3.6b, this can be either a force of attraction or a force of 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 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 9.8 N vertically downwards at every point in the room.

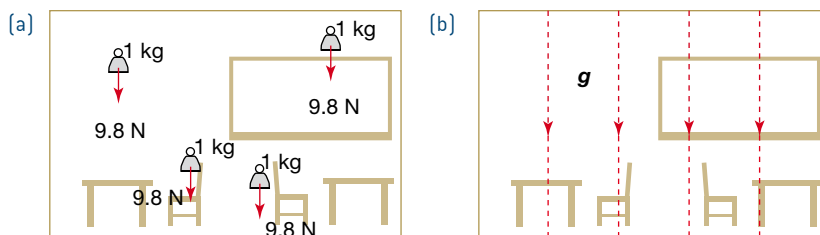


Figure 3.7 (a) The gravitational force on a 1 kg mass is constant at 9.8 N everywhere in your classroom. (b) The uniform gravitational field g is represented by evenly spaced parallel lines in the direction of the force.

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 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 weight of the 1 kg mass decreases as you move further from the centre of the Earth.

At the Earth's surface, a force of 9.8 N acts on the mass. At an altitude of 1000 km, the force-meter would read 7.3 N, and 5000 km up it would show just 3.1 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.

Physics file

Black holes are created when large stars collapse in on themselves at the end of their life cycle. This creates a region where gravity is so strong that light and other forms of radiation cannot escape from the star. Astronomers are only able to detect black holes by studying the behaviour of matter and other stars in the vicinity. A massive black hole, weighing 4 million times as much as the Sun, has been detected at the centre of our Milky Way galaxy.

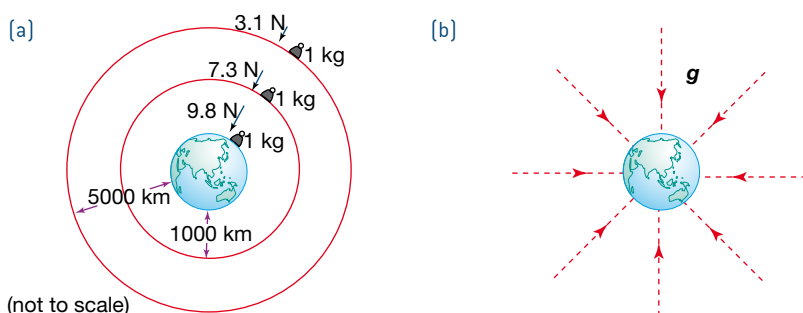


Figure 3.8 (a) The gravitational force acting on a 1 kg mass becomes smaller at greater distances from the Earth. (b) This non-uniform gravitational field, g , is represented by radial lines directed towards the centre of the Earth. The field is strongest where the lines are closest.

The field lines in Figure 3.8b 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.

Calculating the gravitational field strength

The magnitude of a gravitational field is known as the *gravitational field strength*, g . This is defined as the gravitational force that acts on each kilogram of a body in the field. Since gravitational field strength, g , is defined as gravitational force per unit mass, we can say that:

$$g = \frac{F}{m} = \frac{GMm}{R^2} \times \frac{1}{m} = \frac{GM}{R^2}$$



The **GRAVITATIONAL FIELD STRENGTH** is given by:

$$g = \frac{GM}{R^2}$$

If g is known, the weight, F_g or W , of a mass in the field (i.e. the gravitational force acting on the mass) can be found by using:



The **WEIGHT** of a body is given by:

$$F_g = mg = \frac{GMm}{R^2}$$

where G = universal gravitation constant = $6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$

M = central mass [kg]

R = distance from the centre of the central mass [m]

g = gravitational field strength at that location [N kg^{-1}]

The gravitational field strength of the Earth is 9.8 N kg^{-1} 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 9.8 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 (see section 3.4).

Thus, the gravitational field strength at the surface of a central body depends directly on the mass of the body, and is related in an inverse square

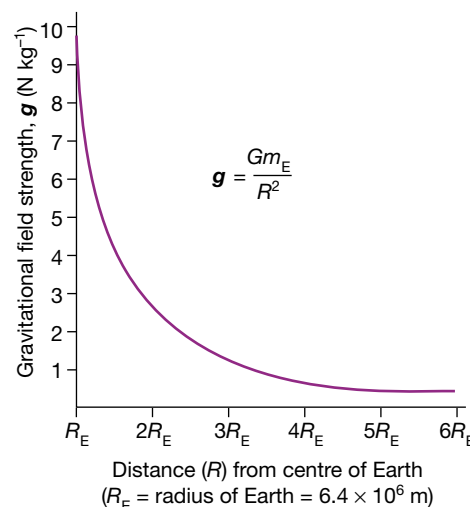


Figure 3.9 The gravitational field strength around the Earth. There is an inverse square relationship between the field strength, g , and the distance, R , from the centre of the Earth.

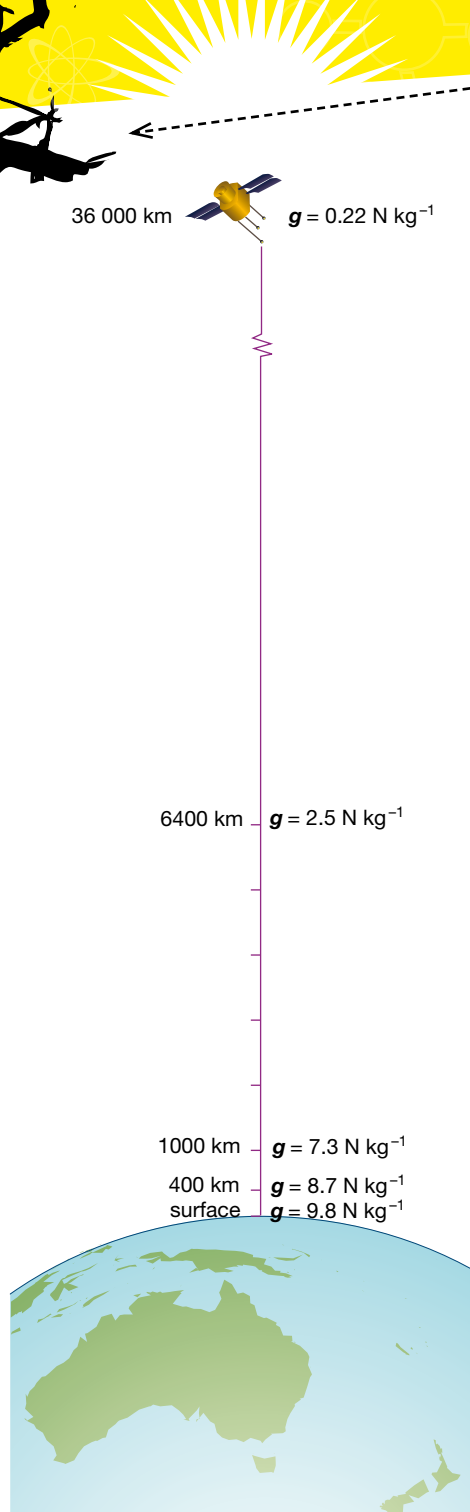


Figure 3.10 At 400 km above the Earth's surface, the gravitational field strength is 8.7 N kg^{-1} . The International Space Station has been in orbit at nearly this altitude since 1998. Australia's Optus communications satellites are in orbit at an altitude of about 36 000 km where Earth's gravitational field has a strength of only 0.22 N kg^{-1} .

manner to the distance from the centre of the body. Some implications of this relationship are as follows.

- The gravitational field of the Earth is actually slightly stronger at the poles, where g is 9.83 N kg^{-1} , than at the Equator, where g is 9.78 N kg^{-1} . This is because the Earth is not perfectly spherical; it is slightly flattened at the poles. The poles are slightly closer to the centre of the Earth.
- The Moon has a gravitational field, but it is only 1.6 N kg^{-1} at its surface. This is much weaker than the Earth's field because the mass of the Moon is about one-eightieth that of the Earth. Counteracting this effect is the fact that the radius is about one-quarter that of the Earth.
- The gravitational fields around asteroids are extremely weak. Again, this is due to their small mass. The largest asteroid, Ceres, has a mass that is just one-fifth that of the Moon. Its field strength is only 0.3 N kg^{-1} at its surface. If you were standing on Ceres and you tossed a rock upwards at 20 m s^{-1} , it would reach a height of over 600 m!

Gravitational fields and freely falling objects

A *freely falling* object is one that is influenced only by a gravitational force. Objects whose motion is significantly affected by air resistance are not considered to be freely falling. Similarly, objects that have an external force such as that provided by a rocket are not in free-fall.

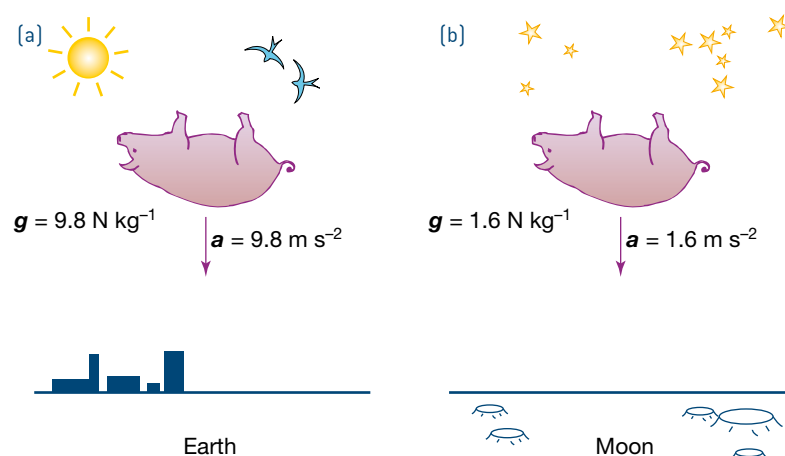


Figure 3.11 The Moon has a weaker gravitational field than the Earth and so freely falling bodies accelerate less than they do on Earth. At each location, the acceleration of a freely falling body is equal to the gravitational field strength.

The net force acting on an object of mass, m , in free-fall is therefore $\Sigma F = mg$.

The acceleration of the object is then $a = \frac{\Sigma F}{m} = \frac{mg}{m} = g$.

In other words, the *acceleration of a freely falling object is equal to the gravitational field strength*. Near Earth, where the gravitational field strength is 9.8 N kg^{-1} , freely falling objects accelerate at 9.8 m s^{-2} . On the Moon, where the strength of gravity is only 1.6 N kg^{-1} , freely falling objects accelerate at 1.6 m s^{-2} .



The **ACCELERATION OF A FREELY FALLING OBJECT** is given by the gravitational field strength, g , at that location.

Table 3.1 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 3.2A

- Calculate the gravitational field strength, g , at the surface of the Earth (mass of Earth is 6.0×10^{24} kg, radius of Earth is 6.4×10^6 m).
- Calculate the gravitational field strength, g , at the surface of the Moon (mass of Moon is 7.34×10^{22} kg, radius of Moon is 1.74×10^6 m).
- Determine the weight of an astronaut, whose total mass is 100 kg, at each of these locations.

Solution

- a At the Earth's surface:

$$\begin{aligned}
 g &= \frac{GM}{R^2} \\
 &= \frac{6.67 \times 10^{-11} \times 6.0 \times 10^{24}}{(6.4 \times 10^6)^2} \\
 &= 9.8 \text{ N kg}^{-1}
 \end{aligned}$$

- b At the Moon's surface:

$$\begin{aligned}
 g &= \frac{GM}{R^2} \\
 &= \frac{6.67 \times 10^{-11} \times 7.34 \times 10^{22}}{(1.74 \times 10^6)^2} \\
 &= 1.62 \text{ N kg}^{-1}
 \end{aligned}$$

- c On Earth, the astronaut would weigh:

$$\begin{aligned}
 F_g &= mg \\
 &= 100 \times 9.8 \\
 &= 980 \text{ N}
 \end{aligned}$$

On the Moon, the astronaut would weigh only:

$$\begin{aligned}
 F_g &= mg \\
 &= 100 \times 1.7 \\
 &= 170 \text{ N}
 \end{aligned}$$

Physics file

It is a simple matter to show that N kg^{-1} and m s^{-2} are equivalent units.

From Newton's second law, $F = ma$, you will remember that:

$$\begin{aligned}
 1 \text{ N} &= 1 \text{ kg m s}^{-2} \\
 \therefore 1 \text{ N kg}^{-1} &= 1 \text{ kg m s}^{-2} \times \text{kg}^{-1} \\
 &= 1 \text{ m s}^{-2}
 \end{aligned}$$

Physics file

Resource companies use sensitive gravimeters (gravimeters) in their search for mineral deposits. The gravitational field above a large deposit of oil is weaker 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 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!



PRACTICAL ACTIVITY 15

Acceleration due to gravity

Worked example 3.2B

A 50 kg piece of space junk is falling towards Earth from a height of 1000 km above the Earth's surface (mass of Earth is 6.0×10^{24} kg, radius of Earth is 6.4×10^6 m). Ignoring air resistance, determine:

- a the gravitational field strength at this height
- b the acceleration of the space junk at this height as it falls
- c the value of the ratio:
$$\frac{\text{field strength at 500 km altitude}}{\text{field strength at 1000 km altitude}}$$

Solution

- a At a height of 1000 km, the distance of the junk from the Earth's centre is:

$$R = 6.4 \times 10^6 + 1.0 \times 10^6 = 7.4 \times 10^6 \text{ m}$$

$$\begin{aligned}\therefore g &= \frac{GM}{R^2} \\ &= \frac{6.67 \times 10^{-11} \times 6.0 \times 10^{24}}{(7.4 \times 10^6)^2} \\ &= 7.3 \text{ N kg}^{-1}\end{aligned}$$

- b The acceleration of the space junk will equal the gravitational field strength, so it will fall with an acceleration of 7.3 m s^{-2} towards Earth.

- c You would expect the gravitational field to be stronger at the lower altitude. This can be confirmed by using a ratio approach.

At 500 km, $R = 6.9 \times 10^6$ m; at 1000 km, $R = 7.4 \times 10^6$ m.

$$\begin{aligned}\frac{g(500 \text{ km})}{g(1000 \text{ km})} &= \frac{GM/R_{500 \text{ km}}^2}{GM/R_{1000 \text{ km}}^2} \\ &= \frac{R_{1000 \text{ km}}^2}{R_{500 \text{ km}}^2} \\ &= \frac{(7.4 \times 10^6)^2}{(6.9 \times 10^6)^2} \\ &= 1.2\end{aligned}$$

The field strength at 500 km altitude is 1.2 times greater than the field strength at 1000 km.



3.2 summary

Gravitational fields

- A gravitational field is a region in which any body will experience a gravitational force.
- The gravitational field strength, g , is greater around more massive central objects, and becomes weaker at greater distances from the central body.
- Gravitational field strength, g , is given by:
$$g = \frac{GM}{R^2}$$
- At the surface of the Earth, the gravitational field strength is 9.8 N kg^{-1} .

- The gravitational force acting on an object in a gravitational field is also called the weight, F_g or W , of the object and is given by:

$$F_g = mg = \frac{GMm}{R^2}$$

- Any object that is falling freely through a gravitational field will fall with an acceleration equal to the gravitational field strength at that location.



3.2 questions

Gravitational fields

For these questions, assume that $G = 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$ and that the gravitational field strength on the surface of the Earth, g , is 9.8 N kg^{-1} .

The following information applies to questions 1–3.

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

- Calculate the gravitational field strength, g , at the surface of each planet.
- Using your answers to Question 1, calculate the weight of an 80 kg astronaut on the surface of:
 - Mercury
 - Saturn
 - Jupiter.
- The result of your calculation for Question 1 should indicate that the gravitational field strength for Saturn is very close in value to that on Earth, i.e. approximately 10 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 both planets have similar gravitational field strengths?

The following information applies to questions 4–6.

A 100 kg meteor is falling towards the Earth from a distance of 4.0 Earth radii from the centre of the Earth ($4.0R_E$).

- Calculate g at this height.
- What is the acceleration of the meteor at this height?
- For this meteor determine the ratio:

$$\frac{\text{acceleration at } 4.0R_E}{\text{acceleration at } 2.0R_E}$$

The following information applies to questions 7–9.

There are bodies outside our Solar System, such as neutron stars, that produce very large gravitational fields. A typical neutron star can have a mass of $3.0 \times 10^{30} \text{ kg}$ and a radius of just 10 km.

- Calculate the gravitational field strength at the surface of such a star.
- Calculate the gravitational field strength at a distance of 5000 km from this star.
- What would be the magnitude of the acceleration of a 10000-tonne asteroid located at this distance and falling towards this star?
- A gravimeter is a device that can measure the Earth's gravitational field strength very accurately. Briefly explain how such a meter could be used to locate mineral deposits.
- Two meteors, X and Y, are falling towards the Moon. Both are $3.0 \times 10^6 \text{ m}$ from the centre of the Moon. Meteor X has a mass of 500 kg and meteor Y has a mass of 50 kg. The mass of the Moon is $7.3 \times 10^{22} \text{ kg}$. Calculate the:
 - gravitational force acting on X
 - acceleration of X
 - gravitational force acting on Y
 - acceleration of Y
 - gravitational field strength at this location.
- 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 $6.0 \times 10^{24} \text{ kg}$, the mass of the Moon is $7.3 \times 10^{22} \text{ kg}$ and the radius of the Moon's orbit is $3.8 \times 10^8 \text{ m}$, calculate the distance of this point from the centre of the Earth.

3.3

Satellites in orbit

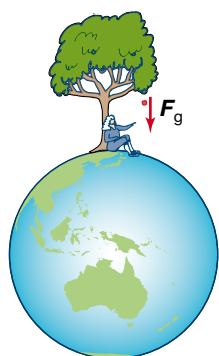


Figure 3.12 Newton realised that the gravitational attraction of the Earth was determining the motion of both the Moon and the apple.

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.

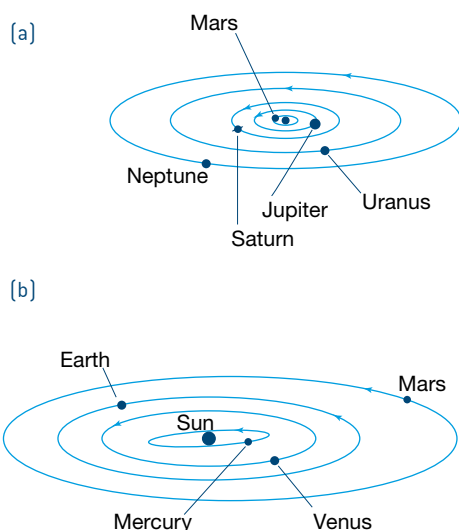


Figure 3.14 The outer planets of the Solar System have such large orbits that they cannot be shown on the same scale diagram as the inner planets. (a) The outer, or jovian, planets are spread apart. If you travelled from the Sun to Neptune, by the time you reached Uranus you would be just one-sixth of 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.

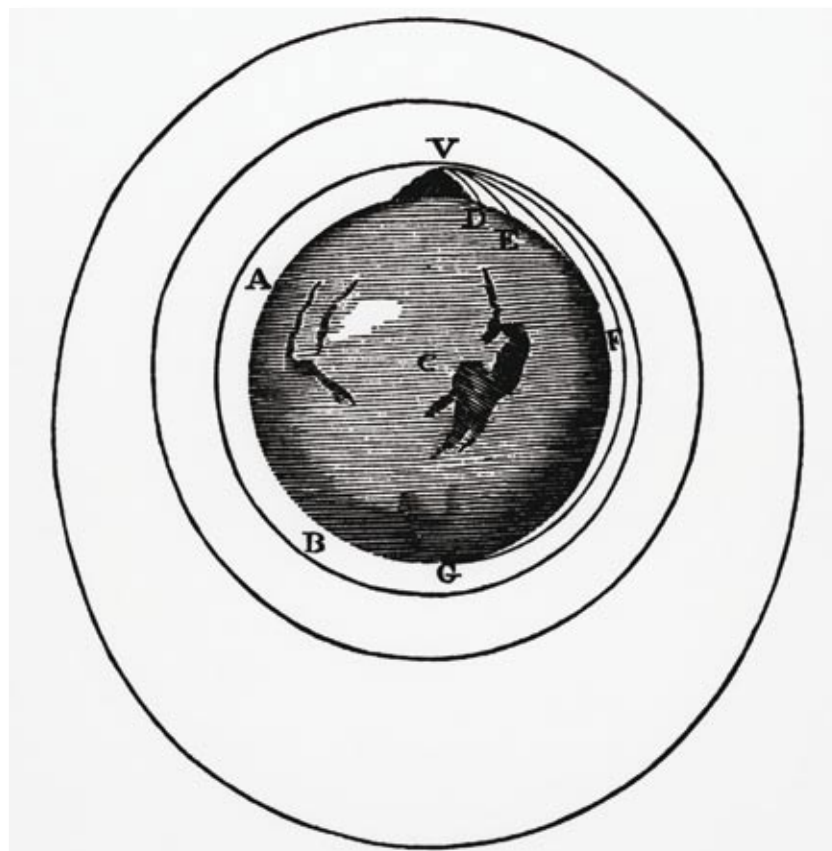


Figure 3.13 Newton's original sketch shows how a projectile that was fired fast enough would fall all the way around the Earth and become an Earth satellite.

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 3.2.

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. Australia uses the Optus satellites for 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 by astronomers 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 and to determine the extent of rainforest destruction. Satellites are playing an important role in monitoring the extent of global warming and climate change.



Figure 3.15 Australia's Optus D1 and D2 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.

Table 3.2 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

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.

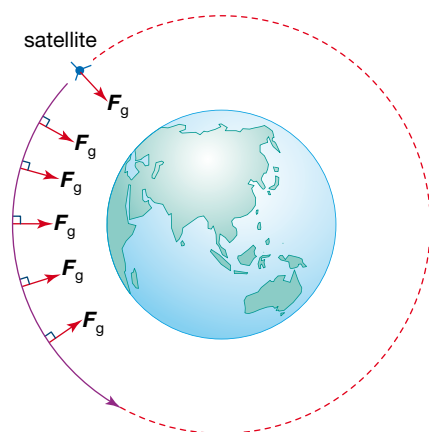


Figure 3.17 The gravitational force that acts on a satellite in a circular orbit is always at right angles to the velocity of the satellite. The force is directed towards the centre of the central mass and gives the satellite a centripetal acceleration.

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.

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.

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.

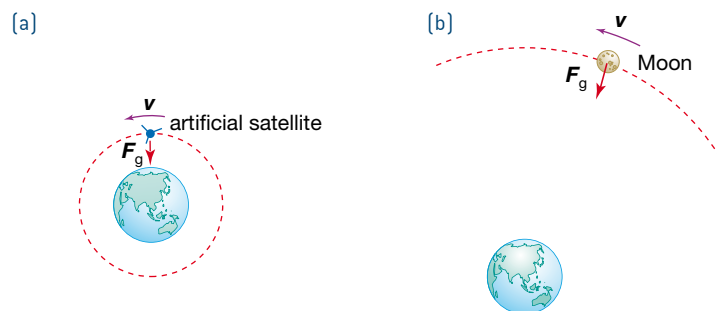


Figure 3.16 The only force acting on (a) an artificial satellite and (b) a natural satellite is the gravitational attraction of the central body.

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 2. 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.

As you will recall from section 3.2, 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 8.2 N kg^{-1} . The circular motion of the HST means that it has a centripetal acceleration of 8.2 m s^{-2} as it orbits.

Table 3.3 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	26 500	20 200	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 3.3 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 2 can be used again. The speed, v , of the satellite can be calculated from its motion for one revolution. It will travel a distance equal to the circumference of the circular orbit ($2\pi R$) in a time T .



The speed of a satellite is given by:

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

The centripetal acceleration, a , 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:

$$\text{Centripetal acceleration } a = \frac{v^2}{R} = \frac{4\pi^2 R}{T^2}$$

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:

$$a = \frac{v^2}{R} = \frac{4\pi^2 R}{T^2} = \frac{GM}{R^2} = g$$

The gravitational force, F_g , acting on the satellite can then be found by using Newton's second law.



GRAVITATIONAL FORCE is given by:

$$F_g = \frac{mv^2}{R} = \frac{4\pi^2 Rm}{T^2} = \frac{GMm}{R^2} = mg$$

where v = speed (m s^{-1})

R = radius of orbit (m)

T = period of orbit (s)

M = central mass (kg)

m = mass of satellite (kg)

g = gravitational field strength at R (N kg^{-1})

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 freely falling objects at the Earth's surface, the *mass* of the satellite itself has *no bearing* on any of these quantities.

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.

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* 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 3.18 Edwin Hubble [1889–1953], after whom the Hubble Space Telescope was named.

Worked example 3.3A

Optus D2 is a geostationary satellite. Its period of orbit is 24 hours so that it revolves at the same rate at which the Earth turns. Given that the mass of the Earth is 6.0×10^{24} kg and the mass of Optus D2 is 1160 kg, calculate:

- Optus D2's orbital radius
- the gravitational field strength at this radius
- Optus D2's orbital speed
- Optus D2's acceleration.

Solution

- Period $T = 24$ hours
 $= 24 \times 60 \times 60$
 $= 86\,400$ s

To calculate R , use:

$$\frac{GM}{R^2} = \frac{4\pi^2 R}{T^2}$$

$$\Rightarrow R^3 = \frac{GMT^2}{4\pi^2}$$

$$\Rightarrow R = \sqrt[3]{\frac{GMT^2}{4\pi^2}}$$

$$= 4.23 \times 10^7 \text{ m}$$

So the orbital radius of Optus D2 is 42 300 km. This is about 5.5 Earth radii from the surface of the Earth.

- $g = \frac{GM}{R^2}$
 $= \frac{6.67 \times 10^{-11} \times 6.0 \times 10^{24}}{(4.23 \times 10^7)^2}$
 $= 0.22 \text{ N kg}^{-1}$

This satellite is in orbit at a great distance from the Earth and the gravitational field strength at its location is just 0.22 N kg^{-1} .

- The speed of Optus D2 is given by:

$$v = \frac{2\pi R}{T}$$

$$= \frac{2\pi \times 4.23 \times 10^7}{86\,400}$$

$$= 3.08 \times 10^3 \text{ m s}^{-1}$$

i.e. about 3 km s^{-1}

- 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. $a = g = 0.22 \text{ m s}^{-2}$ towards the centre of the Earth

Worked example 3.3B

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.

- Use this information to determine the mass of Jupiter.
- Calculate the orbital speed of Ganymede.
- Calculate the gravitational force that Ganymede exerts on Jupiter.
- What is the size of the gravitational force that Jupiter exerts on Ganymede?

Solution

- a To calculate the mass of Jupiter, use:

$$\begin{aligned}\frac{GM}{R^2} &= \frac{4\pi^2 R}{T^2} \\ \Rightarrow M &= \frac{4\pi^2 R^3}{GT^2} \\ &= \frac{4 \times \pi^2 \times (1.07 \times 10^9)^3}{6.67 \times 10^{-11} \times (6.18 \times 10^5)^2} \\ &= 1.90 \times 10^{27} \text{ kg}\end{aligned}$$

- b The orbital speed of Ganymede is:

$$\begin{aligned}v &= \frac{2\pi R}{T} \\ &= \frac{2 \times \pi \times 1.07 \times 10^9}{6.18 \times 10^5} \\ &= 1.09 \times 10^4 \text{ m s}^{-1}, \text{ i.e. about } 11 \text{ km s}^{-1}\end{aligned}$$

- c The gravitational force that Ganymede exerts on Jupiter is:

$$\begin{aligned}F &= \frac{GMm}{R^2} \\ &= \frac{6.67 \times 10^{-11} \times 1.90 \times 10^{27} \times 1.66 \times 10^{23}}{(1.07 \times 10^9)^2} \\ &= 1.84 \times 10^{22} \text{ N}\end{aligned}$$

- d The gravitational force that Jupiter exerts on Ganymede will also equal $1.84 \times 10^{22} \text{ N}$.

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. The Canadian scientists have called it the Humble Space Telescope.

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:

$$F = \frac{GMm}{R^2} = \frac{4\pi^2 Rm}{T^2}$$

$$\Rightarrow \frac{GM}{4\pi^2} = \frac{R^3}{T^2}$$

That is, for a given central body of mass M , the ratio $\frac{R^3}{T^2}$ is constant and equal to $\frac{GM}{4\pi^2}$ 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' .

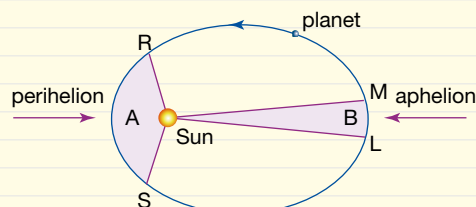


Figure 3.19 Planets orbit in elliptical paths with the Sun at one focus. Their speed varies continually, and they are fastest when closest to the Sun. A line joining a planet to the Sun will sweep out equal areas in equal times, e.g. the time it takes to move from R to S is equal to the time it takes 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 amongst 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 3.4 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 2006 and orbits at 35800 km directly over the Equator. Its closest point to the Earth, or perigee, is 35776 km. Its furthest point from the Earth is known as the apogee at 35798 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. Hubble 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. The HST is in a low-Earth orbit inclined at 28° to the Equator. Its expected life span was about 15 years, but a service and repair mission scheduled for 2008 will, if successful, extend its life by another 5 years.

Table 3.5 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
Hubble	Inclined	28°	591	599	96.6 min
NOAA-18	Polar	99°	846	866	102 min

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 3.20 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.



3.3 summary

Satellites in orbit

- 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:
- The acceleration of a satellite in a circular orbit is given by:

$$a = \frac{v^2}{R} = \frac{4\pi^2 R}{T^2} = \frac{GM}{R^2} = g$$

- The gravitational force acting on a satellite in a circular orbit is given by:

$$F_g = \frac{mv^2}{R} = \frac{4\pi^2 Rm}{T^2} = \frac{GMm}{R^2} = mg$$

- For any central body of mass M :

$$\frac{R^3}{T^2} = \frac{GM}{4\pi^2} = \text{constant for all satellites of this body}$$

So knowing another satellite's orbital radius, R , enables its period, T , to be determined.

$$v = \frac{2\pi R}{T}$$



3.3 questions

Satellites in orbit

1 Which of the following statements is correct? A satellite in a stable circular orbit 100 km above the Earth will move:

- A with an acceleration of 9.8 m s^{-2}
- B with a constant velocity
- C with zero acceleration
- D with 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 0.22 N kg^{-1} . The mass of this satellite is $2.3 \times 10^3 \text{ kg}$.

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 $1.02 \times 10^{26} \text{ kg}$. One of its moons, Triton, has a mass of $2.14 \times 10^{22} \text{ kg}$ and an orbital radius equal to $3.55 \times 10^8 \text{ m}$. 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 $1.35 \times 10^{23} \text{ kg}$ and an orbital radius of $1.22 \times 10^9 \text{ m}$. The orbital period of Titan is $1.38 \times 10^6 \text{ s}$.

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 $4.87 \times 10^{24} \text{ kg}$ and a radius of $6.05 \times 10^6 \text{ m}$. The length of a day on Venus is $2.10 \times 10^7 \text{ s}$. 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.

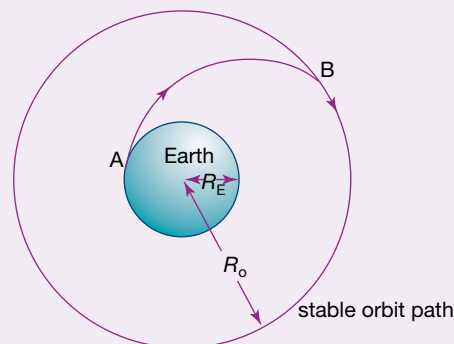
	Orbital radius (m)	Orbital period (days)
Atlas	1.37×10^8	0.602
Helene	3.77×10^8	2.75

a Calculate the value of these ratios:

- i orbital speed of Atlas/orbital speed of Helene
- ii acceleration of Atlas/acceleration of Helene.

b The largest of Saturn's moons is Titan. It has an orbital radius of $1.20 \times 10^9 \text{ m}$. 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 $6.4 \times 10^6 \text{ m}$. The ratio of the gravitational field strength at A to that at B is equal to 1.2. Calculate the distance R_o .



11 Melbourne TV stations can use pictures from the Japanese MTSAT-1R weather satellite, which is in a geostationary orbit over the Equator to the north of Australia. Why is it not possible to place a satellite in orbit so that it is always directly over Melbourne?

12 Ceres, the first asteroid to be discovered, was found by Giuseppe Piazzi in 1801. Ceres has mass $7.0 \times 10^{20} \text{ kg}$ and radius 385 km.

- a What is the gravitational field strength at the surface of Ceres?
- b Determine the speed required by a satellite in order to remain in orbit 10 km above the surface of Ceres.

3.4 Energy changes in gravitational fields

Consider the motion of an unpowered projectile such as a rock, moving freely through the gravitational field of the Earth. The gravitational field of the Earth does work on the projectile as it moves through the field, changing its kinetic energy (ΔE_k) with an equal but opposite change in gravitational potential energy (ΔU_g).

In accordance with the principle of conservation of energy, the total energy of any projectile in free-fall remains constant. For motion near the Earth's surface, the gravitational field strength is taken to be constant, so the force acting on the rock is also constant. The projectile will therefore fall with uniform acceleration equal to the gravitational field strength, g .

In this section we will analyse the *energy changes* experienced by satellites moving in *circular orbits*, as well as the energy changes of objects such as meteors that travel vast distances through *changing gravitational fields*.

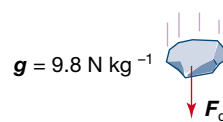


Figure 3.21 The gravitational field does work on the rock and causes it to gain kinetic energy. The increase in kinetic energy is equal to the loss in gravitational potential energy that the rock experiences.

Energy of satellites in circular orbits

A satellite in a circular orbit moves at a constant distance from the centre of a massive body. The *potential energy* of the satellite remains constant. The *kinetic energy* of the satellite also remains constant because the gravitational force acting on it is always perpendicular to its motion. This means that the *gravitational force does no work* on the satellite (see section 3.3) and so does not cause any change in its total energy.

The gravitational force that acts on the satellite is directed towards the central body and is constant. So the satellite will have uniform acceleration equal to the gravitational field strength, directed towards the central body (see Figure 3.22).

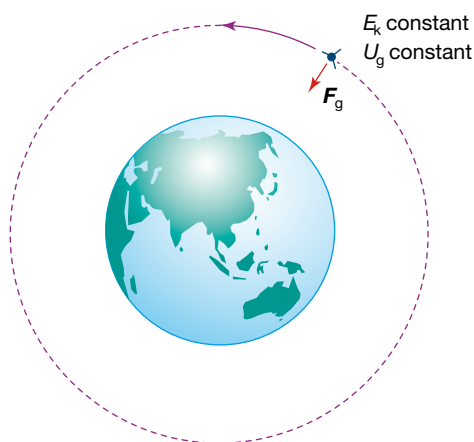


Figure 3.22 No work is done as a satellite completes its orbit, since the force F_g acts only to change the direction of the satellite. Its speed and altitude are unaltered, and so its kinetic energy and gravitational potential energy remain constant.

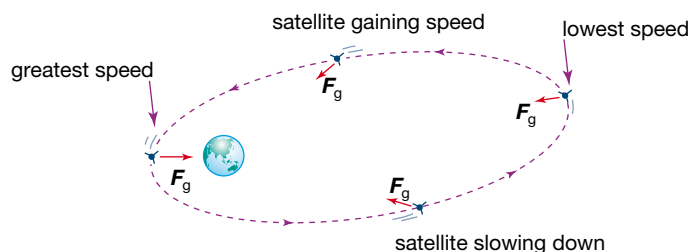


Figure 3.23 The gravitational force is not perpendicular to the velocity of a satellite in an elliptical orbit, and so does work. It acts to slow the satellite as it moves away from the central body, and increase its speed as it approaches. The total energy of the satellite remains constant at all times throughout the orbit.

Physics file

Most natural and artificial satellites move in elliptical orbits. They travel at different distances from the central body and so the gravitational field that they pass through changes in strength. The gravitational force that is acting also varies in strength and is not perpendicular to the motion. The force has a component in the direction of the motion of the satellite, as well as a component perpendicular to the direction of the motion. This causes the speed, kinetic energy and gravitational potential energy of the satellite to change throughout the orbit, as shown in Figure 3.23. The total energy of the satellite, however, remains constant at all times.

Energy of objects moving through a changing gravitational field

Consider the example of a 10 kg meteor falling towards the Earth from deep space. Closer to the Earth, the meteor moves through regions of increasing gravitational field strength.

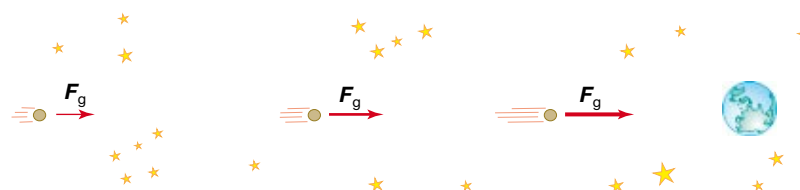


Figure 3.24 As a meteor approaches Earth, it moves through an increasingly stronger gravitational field and so is acted upon by a greater gravitational force. The work done by this force can be determined from a force–distance graph.

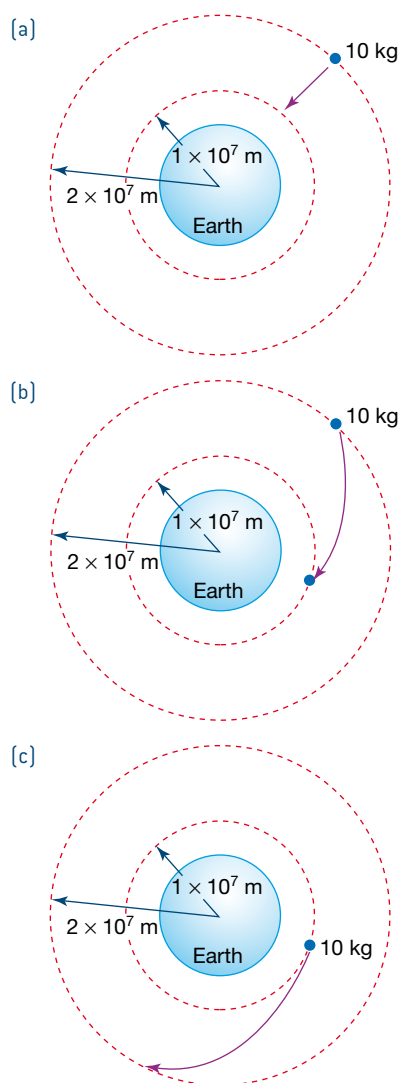


Figure 3.26 The shaded region on the gravitational force–distance graph in Figure 3.24 could represent the changes in kinetic and gravitational potential energy in these free-fall situations. (a) A 10 kg body falling directly towards Earth will gain E_k and lose U_g . (b) A 10 kg body falling obliquely towards Earth will gain E_k and lose U_g . (c) A 10 kg body moving away from Earth will lose E_k and gain U_g .

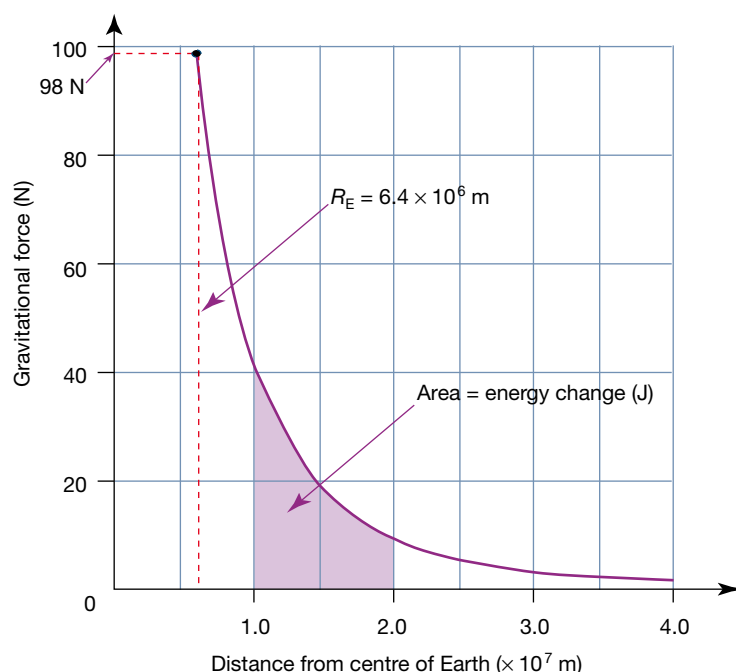


Figure 3.25 The gravitational force acting on a 10 kg body at different distances from the Earth. The shaded region represents the work done by the gravitational field as the body moves between 2.0×10^7 m and 1.0×10^7 m from the centre of the Earth.

When a free-falling body like the meteor is acted upon by a *varying gravitational force*, the energy changes of the body can be analysed by using *gravitational force–distance graphs*. Now, the *area* under a force–distance graph is equal to *work done*, i.e. the *energy change* of a body. The area under the graph has units of newton metres (N m), equivalent to joules (J). The graph in Figure 3.25 enables us to analyse the motion of the meteor.



The **AREA** under a gravitational force–distance graph gives the *change in energy* that an object will experience as it moves freely through the gravitational field.

The gravitational force–distance graph can be used to find both the change in kinetic energy, ΔE_k , and the change in gravitational potential energy, ΔU_g .

Consider, say, a 10 kg meteor as it moves from one point to another. The shaded area in Figure 3.25 represents the increase in kinetic energy of the meteor as it falls from a distance of 2.0×10^7 m to 1.0×10^7 m from the centre of the Earth. This is the same even if the meteor falls obliquely (Figure 3.26b). This area also represents the amount of gravitational potential energy that the meteor loses as it approaches Earth.

An alternative situation could be that the 10 kg mass is an unpowered satellite travelling away from Earth. In this case, the shaded area would give the loss in kinetic energy and the increase in gravitational potential energy of the satellite as it moves from 1.0×10^7 m to 2.0×10^7 m from the Earth's centre (Figure 3.26c).



Work done by gravitational field = area under force–distance (F – d) graph
 $= \Delta E_k = -\Delta U_g$

Energy changes in gravitational fields can also be found from gravitational field–distance graphs (Figure 3.26). The area of a g – d graph gives a quantity that has units of $\text{N kg}^{-1} \times \text{m}$, which is equivalent to J kg^{-1} . So the area indicates the change in energy for each kilogram of the object's mass. To find the work done or energy change (J), the area (J kg^{-1}) must therefore be multiplied by the mass (kg).



The AREA under a gravitational field–distance (g – d) graph indicates the energy change per kilogram of mass. To find the change in energy, the area is multiplied by the mass (kg) of the object.

Gravitational force–distance and field–distance graphs are curved and so the usual method of determining the area under them is by using a *counting squares* technique.

Worked example 3.4A

A 500 kg lump of space junk is plummeting towards the Moon as shown. Its speed when it is 2.7×10^6 m from the centre of the Moon is 250 m s^{-1} . The Moon has a radius of 1.7×10^6 m.

Using the gravitational force–distance graph, determine:

- the initial kinetic energy of the junk
- the increase in kinetic energy of the junk as it falls to the Moon's surface
- the speed of the junk as it crashes into the Moon
- the gravitational field strength at an altitude of 500 km.

Solution

- a** Initial kinetic energy of junk:

$$\begin{aligned} E_k &= \frac{1}{2}mv^2 \\ &= 0.5 \times 500 \times 250^2 \\ &= 1.6 \times 10^7 \text{ J} \end{aligned}$$

- b** To find the increase in kinetic energy of the junk as it falls from a distance of 2.7×10^6 m to the surface, it is necessary to estimate the area under the force–distance graph between 2.7×10^6 m and 1.7×10^6 m from the Moon's centre. There are approximately 53 squares under this part of the graph. Each square represents:

$$100 \text{ N} \times 0.1 \times 10^6 \text{ m} = 1.0 \times 10^7 \text{ J}$$

Thus the increase in kinetic energy of the junk as it falls is:

$$53 \times 1.0 \times 10^7 = 5.3 \times 10^8 \text{ J}$$

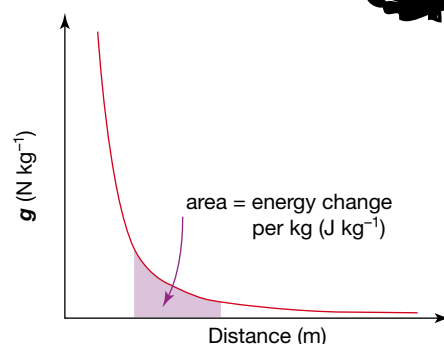
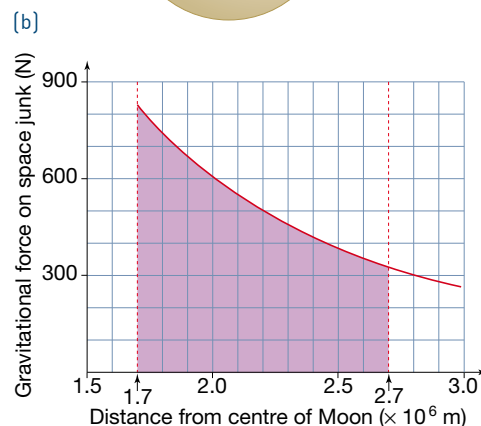
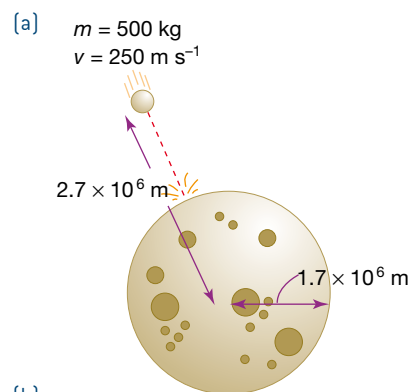


Figure 3.27 Graphs of gravitational field strength versus distance (g vs. R) can also be used to analyse the energy changes of a body moving through a gravitational field.



Physics file

One of the more unusual satellites, SuitSat, was launched from the International Space Station on 3 February 2006. It was an obsolete Russian spacesuit into which the astronauts had placed a radio transmitter, batteries and some sensors. Its launch involved simply being pushed off by one of the astronauts while on a spacewalk. SuitSat was meant to transmit signals that would be picked up by ham radio operators on Earth for a few weeks, but transmissions ceased after just a few hours. It burned up in the atmosphere over Western Australia in September 2006. More SuitSat missions are planned!



Figure 3.28 This photograph is not showing an astronaut drifting off to certain death in space. This is SuitSat—one of the strangest satellites ever launched—at the start of its 7-month mission.

- c** The space junk has 1.6×10^7 J of kinetic energy initially, and gains 5.3×10^8 J as it falls. Therefore, its kinetic energy as it strikes the Moon is:

$$E_k = 1.6 \times 10^7 + 5.3 \times 10^8 \\ = 5.46 \times 10^8 \text{ J}$$

To find the speed of the space junk as it reaches the Moon's surface:

$$E_k = \frac{1}{2}mv^2 \\ = 5.46 \times 10^8 \\ \Rightarrow v = \sqrt{\frac{5.46 \times 10^8}{0.5 \times 500}} \\ = 1480 \text{ m s}^{-1}$$

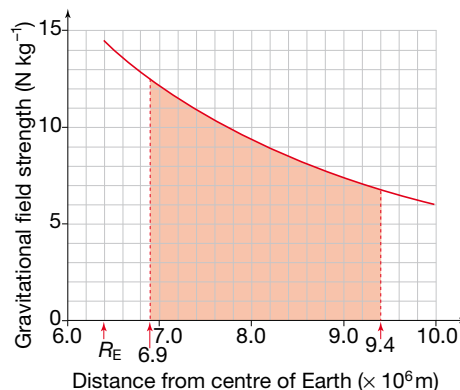
The speed of the space junk on impact with the Moon is $1.5 \times 10^3 \text{ m s}^{-1}$ or 1.5 km s^{-1} .

- d** 500 km is equal to 5.0×10^5 m. This is a distance from the centre of the Moon of $1.7 \times 10^6 + 5.0 \times 10^5 = 2.2 \times 10^6$ m. At this distance, the gravitational force on the 500 kg body is (reading from the graph) equal to 500 N. Thus the gravitational field strength at this altitude is:

$$g = \frac{F}{m} \\ = \frac{500}{500} \\ = 1.0 \text{ N kg}^{-1}$$

Worked example 3.4B

A wayward satellite of mass 1500 kg has developed a highly elliptical orbit around the Earth. At its closest approach (perigee), it is just 500 km above the Earth's surface. Its furthest point (apogee) is 3000 km from the Earth's surface. The Earth has a mass of 6.0×10^{24} kg and a radius of 6.4×10^6 m. The gravitational field strength of the Earth is shown in the graph.



- a** Calculate the change in potential energy of the satellite as it moves from the closest to the furthest point from the Earth.
b The satellite was moving with a speed of 15 km s^{-1} at its closest point to Earth. How fast was it travelling at its furthest point?

Solution

- a** Closest point = $500 \text{ km} + 6.4 \times 10^6 \text{ m}$
 $= 5.0 \times 10^5 + 6.4 \times 10^6 \text{ m}$
 $= 6.9 \times 10^6 \text{ m}$
 Furthest point = $3000 \text{ km} + 6.4 \times 10^6 \text{ m}$
 $= 3.0 \times 10^6 + 6.4 \times 10^6 \text{ m}$
 $= 9.4 \times 10^6 \text{ m}$

The area under the g - d graph between $6.9 \times 10^6 \text{ m}$ and $9.4 \times 10^6 \text{ m}$ must first be determined. This is the area shaded on the graph, which gives the energy per kilogram (J kg^{-1}).

$$\begin{aligned}\text{Area} &= 120 \text{ squares} \times (0.2 \times 10^6) \times 1.0 = 2.4 \times 10^7 \text{ joules per kilogram} \\ \text{Change in energy} &= \text{area under graph} \times \text{mass of satellite} \\ &= 2.4 \times 10^7 \times 1500 \\ &= 3.6 \times 10^{10} \text{ J}\end{aligned}$$

The satellite has moved away from Earth so has gained 3.6×10^{10} J of potential energy.

b Its kinetic energy at the closest point is:

$$\begin{aligned}E_k &= \frac{1}{2}mv^2 \\ &= 0.5 \times 1500 \times (1.5 \times 10^4)^2 \\ &= 1.7 \times 10^{11} \text{ J}\end{aligned}$$

From part a, the satellite must lose 3.6×10^{10} J of kinetic energy as it travels away from perigee to apogee.

$$E_k (\text{apogee}) = 1.7 \times 10^{11} - 3.6 \times 10^{10} = 1.3 \times 10^{11} \text{ J}$$

$$\frac{1}{2}mv^2 = 1.3 \times 10^{11}$$

$$\begin{aligned}v &= \sqrt{\frac{2 \times 1.3 \times 10^{11}}{1500}} \\ &= 1.3 \times 10^4 \text{ m s}^{-1} \\ &= 13 \text{ km s}^{-1}\end{aligned}$$



3.4 summary

Energy changes in gravitational fields

- The total energy of a body moving freely through a gravitational field is constant, although the relative amounts of kinetic energy and gravitational potential energy may change.
- A satellite in a stable circular orbit has a constant amount of both kinetic energy and gravitational potential energy.
- The energy changes of a body moving freely through a gravitational field can be determined from the gravitational force–distance graph for that body.
- The area under a gravitational force–distance graph gives the change in kinetic energy, ΔE_k , or change in gravitational potential energy, ΔU_g , of a body, and indicates the work done by the gravitational field.
- The area under a gravitational field–distance graph gives the change in energy per kilogram (J kg^{-1}) of the object. To calculate the energy change, the area is multiplied by the mass (kg).



3.4 questions

Energy changes in gravitational fields

In these questions, the following data may be used: radius of Earth = 6.4×10^6 m, mass of Earth = 6.0×10^{24} kg.

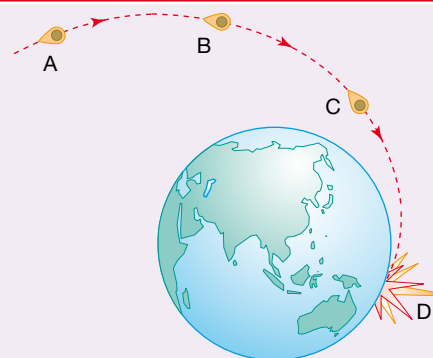
1 Which one of the following statements is correct?

A satellite in a stable circular orbit around the Earth will have:

- A** varying potential energy as it orbits
- B** varying kinetic energy as it orbits
- C** constant kinetic energy and constant potential energy.

The following information applies to questions 2–4.

Ignore air resistance when answering these questions. The path of a meteor plunging towards the Earth is as shown.



- 2** How does the gravitational field strength change as the meteor travels from point A to point D?
- 3** Discuss any changes in the acceleration of the meteor as it travels along the path shown.

Physics file

Students of mathematics will appreciate that the area under the force–distance graph can also be determined by using calculus:

ΔE as a body moves between altitudes A and B

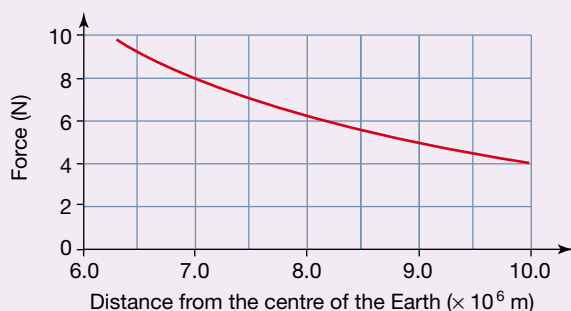
$$= \text{area} = \int_A^B F \cdot dR = \int_A^B \frac{GMm}{R^2} dR$$

$$= \left[\frac{-GMm}{R} \right]_A^B$$

4 Which one or more of the following statements is correct?

- A The kinetic energy of the meteor increases as it travels from A to D.
- B The gravitational potential energy of the meteor decreases as it travels from A to D.
- C The total energy of the meteor remains constant.
- D The total energy of the meteor increases.

The following information applies to questions 5 and 6. The graph shows the force on a mass of 1.0 kg as a function of its distance from the centre of the Earth.



- 5 a Use the graph to determine the gravitational force between the Earth and a 1.0 kg mass 100 km above the Earth's surface.
- b Use the graph to determine the height above the Earth's surface at which a 1.0 kg mass would experience a gravitational force of 5.0 N.
- 6 A meteorite of mass 1.0 kg is speeding towards the Earth. When this meteorite is at a distance of 9.5×10^6 m from the centre of the planet, its speed is 4.0 km s^{-1} .
- a Determine the kinetic energy of the meteorite when it is 9.5×10^6 m from the centre of the Earth.
 - b Calculate the increase in kinetic energy of the meteorite as it moves from a distance of 9.5×10^6 m from the centre of the Earth to a point that is 6.5×10^6 m from the centre.
 - c Ignoring air resistance, what is the kinetic energy of the meteorite when it is 6.5×10^6 m from the centre of the Earth?

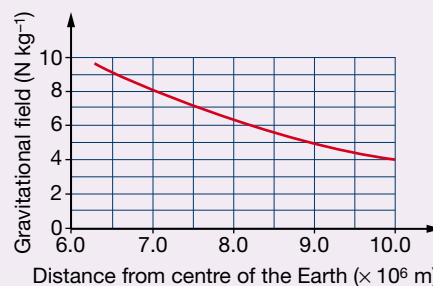
d How fast is the meteorite travelling when it is 6.5×10^6 m from the centre of Earth?

The following information applies to questions 7–9. A 20-tonne remote sensing satellite is in a circular orbit around the Earth at an altitude of 600 km.

- 7 Calculate:
- a the orbital speed of the satellite
 - b the kinetic energy of the satellite
 - c the orbital period of the satellite (in seconds).

8 Ground control decides that this orbit is unsuitable and gives the signal for the satellite's booster rockets to fire. The satellite now moves to a new circular orbit with an altitude of 2600 km. For this new orbit, calculate:

- a the orbital speed of the satellite
 - b the kinetic energy of the satellite
 - c the orbital period of the satellite (in seconds).
- 9 Use the following graph to estimate the increase in gravitational potential energy of the satellite as it moved from its lower orbit to its higher orbit.



- 10 A communications satellite of mass 240 kg is launched from a space shuttle that is in orbit 600 km above the Earth's surface. The satellite travels directly away from the Earth and reaches a maximum distance of 8000 km from the centre of the Earth before stopping due to the influence of the Earth's gravitational field. Use the graph in question 5 to estimate the kinetic energy of this satellite as it was launched.

3.5 Apparent weight and weightlessness

In everyday life, *mass* and *weight* are often given the same meaning. However, mass and weight are *not* the same thing. The mass of a given object will be the same in any location or situation. For example, a 3.0 kg cat will still have a mass of 3.0 kg if it is placed on the Moon; and even if it is drifting around in deep space where the strength of gravity is zero, its mass will still be 3.0 kg.

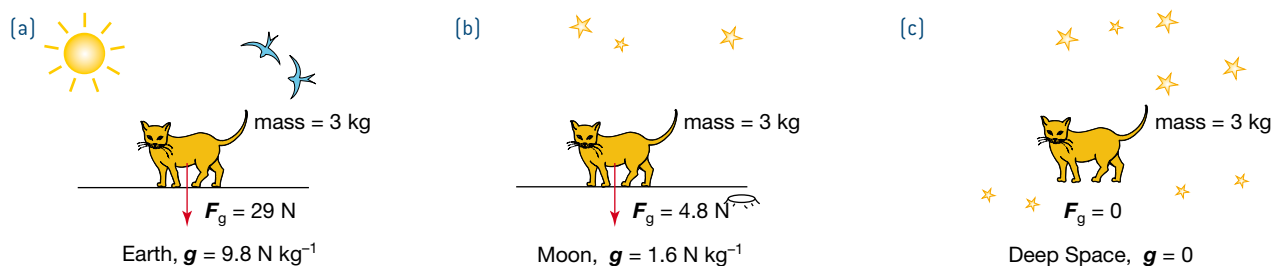


Figure 3.29 The mass of the cat [its resistance to acceleration] is unaffected by the strength of the gravitational field.

To see that the mass of the cat is constant, consider the effect of applying an unbalanced force of 3 N to the cat in the situations above. At each of the locations shown in Figure 3.29, if friction were negligible, the cat would accelerate at 1 m s^{-2} . That is, its resistance to motion is not affected by its location. The mass of the cat determines the acceleration resulting from a given force: $\mathbf{a} = \frac{\mathbf{F}}{m}$.



MASS is a measure of a body's inertia, i.e. its resistance to acceleration.

On the other hand, weight, \mathbf{F}_g or \mathbf{W} , is the *force of gravitational attraction*. Weight is a force and so is measured in newtons (N).

Consider again the 3 kg cat in different locations. Its weight on Earth, where the strength of gravity is 9.8 N kg^{-1} , is:

$$\mathbf{F}_g = m\mathbf{g} = 3.0 \times 9.8 = 29 \text{ N down}$$

On the Moon, the force of gravity acting on the cat would be less than this. The gravitational field strength on the Moon is 1.6 N kg^{-1} and so the weight of the cat, as shown in Figure 3.30b, would be only:

$$\mathbf{F}_g = m\mathbf{g} = 3.0 \times 1.6 = 4.8 \text{ N down}$$

In deep space, far away from any stars or planets, the gravitational field strength is zero. The gravitational force acting on a cat placed here, as seen in Figure 3.30c, would be zero. The cat would experience *true* weightlessness.

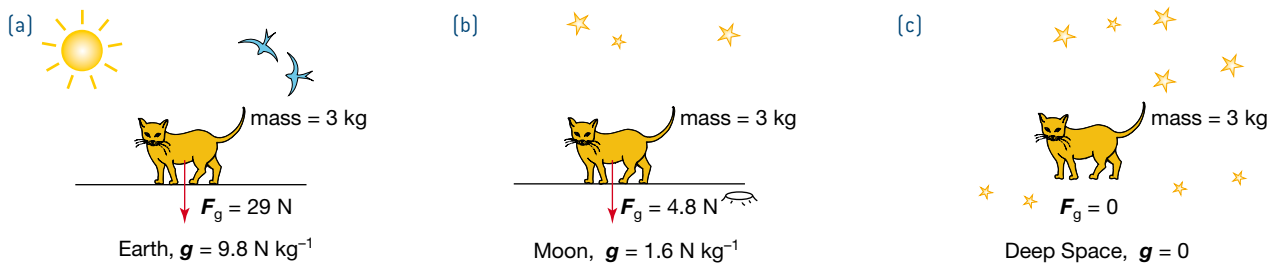


Figure 3.30 The weight of the cat is directly related to the strength of the gravitational field.

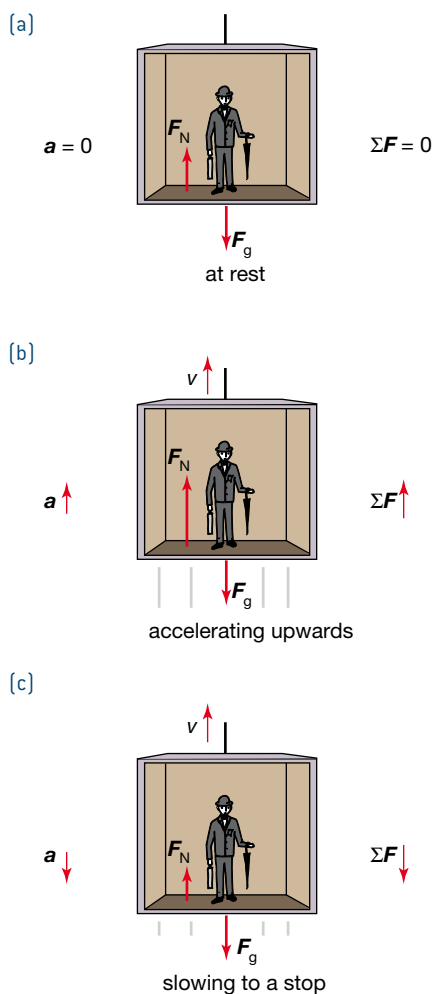
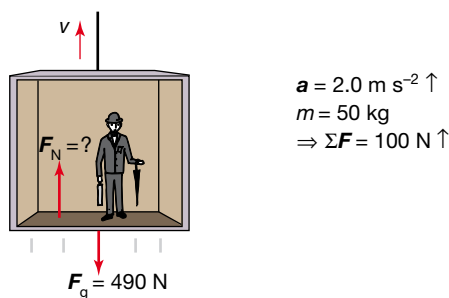


Figure 3.31 The size of the normal force acting on the person in the lift changes throughout the ride.
 (a) When at rest, the magnitude of the normal force, F_N , is equal to the weight of the person, i.e. $F_N = F_g$.
 (b) When accelerating upwards, the normal force, F_N , is greater than the weight of the person, i.e. $F_N > F_g$.
 (c) When slowing to a stop at the top, the normal force, F_N , is less than the weight of the person, i.e. $F_N < F_g$.



Apparent weight

If you have ever travelled in a lift in a city skyscraper, you might have noticed that you feel slightly lighter or slightly heavier on your feet as the lift stops and starts. The motion of the lift seems to make your weight change. You could confirm this sensation by simply standing on a set of bathroom scales in the lift! The reading from the scales would alter throughout the ride.

It is important to keep in mind that your weight has not actually changed at all. Your weight is the force of gravitation that acts on you and this has been constant during your trip to the top of the skyscraper. The quantity that has changed during the lift ride is the *normal force*, F_N or N , that the floor has exerted on your feet.

The net force on a person in a lift waiting for it to start is zero, and so the normal force, F_N , on the person will balance their weight, F_g . This is shown in Figure 3.31a.

As the lift accelerates upwards, a net upwards force acts on the person. The normal force, F_N , from the floor increases and is now greater than the weight, F_g . This is shown in Figure 3.31b. In other words, the floor is pushing harder on the feet and this makes the person feel heavier. This is described as an increase in *apparent weight*.

The opposite happens when the lift slows to stop at the top. As the person slows down, the acceleration is downwards and so the net force experienced is also downwards. As in Figure 3.31c, the normal force, F_N , from the floor is now less than the weight, F_g . The person feels lighter on their feet as they reach the top of the skyscraper. In this case, *apparent weight* has decreased.



The **APPARENT WEIGHT** of a body is equal to the size of the normal force, F_N , that acts on the body.

Worked example 3.5A

James, of mass 50 kg, is in a lift that is accelerating upwards at 2.0 m s^{-2} .

- What is James's mass as the lift accelerates upwards?
- What is James's weight as the lift accelerates upwards?
- Calculate James's apparent weight as the lift accelerates upwards.

Solution

- James's mass is 50 kg.
- James's weight is: $F_g = mg = 50 \times 9.8 = 490 \text{ N}$
- His apparent weight is given by the normal force that the floor exerts on him.
 $a = 2.0 \text{ m s}^{-2} \text{ up}, m = 50 \text{ kg}$
 $\Sigma F = ma = 50 \times 2.0 = 100 \text{ N} \uparrow$
 $F_N + F_g = 100 \text{ N} \uparrow$
 $F_N + 490 \text{ N} \downarrow = 100 \text{ N} \uparrow$
 $F_N = 100 \text{ N} \uparrow - 490 \text{ N} \downarrow = 590 \text{ N} \uparrow$

The normal force that the floor exerts on James is 590 N up. Thus, his apparent weight is 590 N, which is more than his weight of 490 N. James would feel heavier than usual as the lift was accelerating upwards.

Apparent weightlessness and true weightlessness

Astronauts in orbit seem to ‘float’ around as if there were no gravity, yet they are in orbit only a few hundred kilometres above the Earth’s surface where we have established that there are still strong gravitational forces acting. At an altitude of 400 km (a typical shuttle orbit), the gravitational field strength is 8.7 N kg^{-1} , almost as strong as the field strength at the Earth’s surface. Astronauts in orbit do not experience *true weightlessness*. True weightlessness occurs only in deep space, out of reach of the gravitational fields of stars and planets, where a person’s weight is as close to zero as it can be. When there are no gravitational forces acting, objects move as Newton’s first law predicts they should. They either remain at rest or travel in straight lines with constant speed.

To understand *apparent weightlessness*, consider again the case of James, mass 50 kg, the unfortunate occupant of a lift whose cable has snapped. The forces acting on James when he is standing in the lift are his weight, F_g , and the normal force, F_N , from the floor. James is in free-fall and so is accelerating downwards at 9.8 m s^{-2} , the rate determined by the strength of the gravitational field. The net force acting on him is:

$$\Sigma F = ma = 50 \times 9.8 = 490 \text{ N downwards}$$

The gravitational force acting on James is:

$$F_g = mg = 50 \times 9.8 = 490 \text{ N downwards}$$

Looking at these forces, it can be seen that the weight force completely accounts for the net force; the normal force from the floor is zero. The floor exerts no force at all on James. While he is falling, James will be in a state of apparent weightlessness. He will not feel any force from the floor of the lift, and will drift freely inside it while it is plummeting.



Any person or object will experience **APPARENT WEIGHTLESSNESS** when falling with an acceleration equal to the gravitational field strength.

All the people shown in Figure 3.33 are in free-fall and experience zero normal force. They are, for a short time anyway, in a state of apparent weightlessness. Astronauts in an orbiting spacecraft are in a state of continual free-fall since they have an acceleration due to their circular orbit which is equal to the gravitational field strength at their altitude.

At an altitude of 395 km, where the gravitational field strength is 8.7 N kg^{-1} , the astronauts and their spacecraft orbit at a speed that gives them a centripetal acceleration of 8.7 m s^{-2} . This situation is no different from the earlier situation of James falling freely in the lift, when both he and the lift were falling at 9.8 m s^{-2} (the rate determined by the gravitational field). He was in a state of apparent weightlessness. The astronauts and their spacecraft all have an acceleration of 8.7 m s^{-2} (the rate determined by the gravitational field strength at their altitude) towards the Earth. There is no normal force between the astronauts and the floor (or ceiling!) of their spacecraft. They too are in a state of apparent weightlessness as they orbit the Earth. The only difference is that the astronauts’ velocity is directed at 90° to their acceleration.



PRACTICAL ACTIVITY 16

Apparent weight

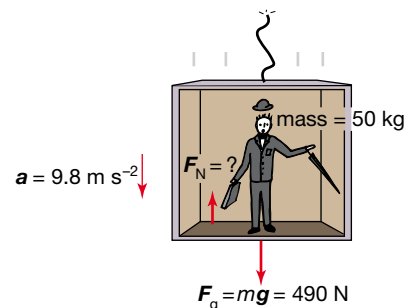


Figure 3.32 The occupant of the plummeting lift has an acceleration equal to that determined by the gravitational field. The normal force is zero, and so he is in a state of apparent weightlessness as he free-falls.

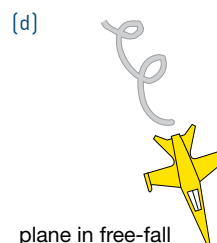
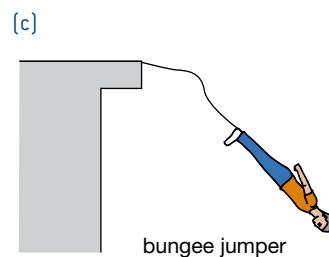
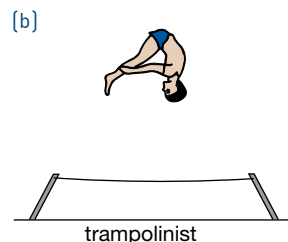
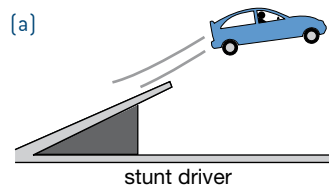


Figure 3.33 All of the participants in these activities accelerate at a rate determined by the gravitational field, and so are in a state of apparent weightlessness.

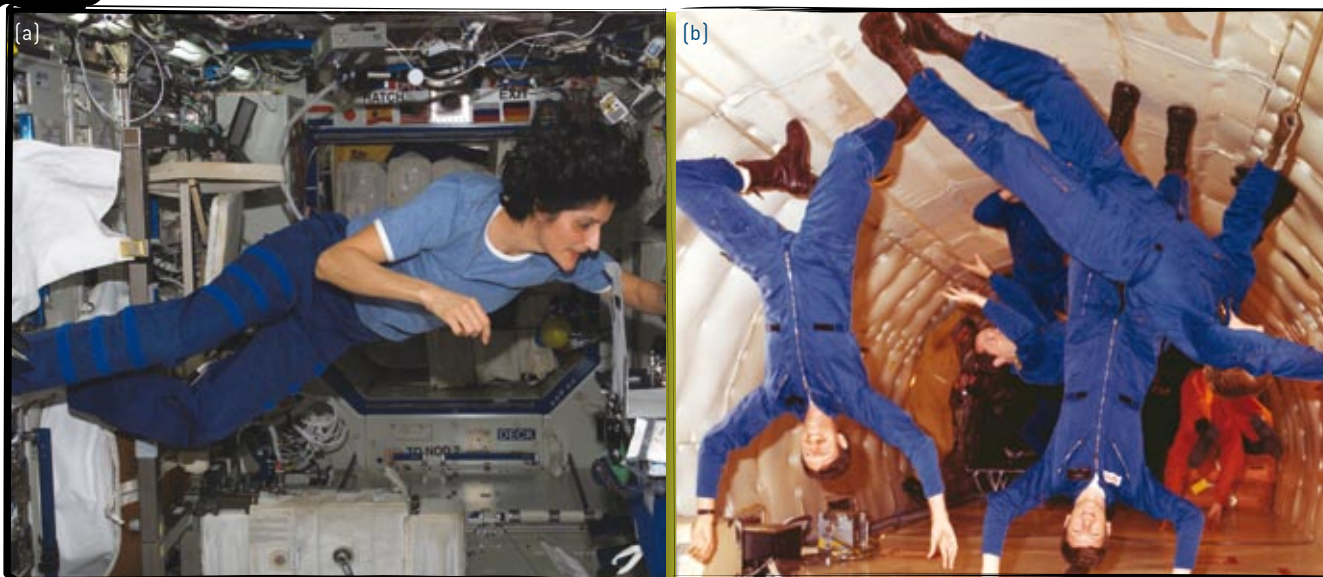


Figure 3.34 (a) This astronaut is in orbit at an altitude of 400 km. The gravitational field strength here is 8.7 N kg^{-1} . She is in free-fall: the only force acting on her is gravity. She moves with an acceleration of 8.7 m s^{-2} towards Earth and is in a continual state of apparent weightlessness. (b) These trainee astronauts are in the 'vomit comet'. The engines of the plane are effectively turned off at an altitude of around 12 km and the plane free-falls in a parabolic arc for about 20 s. The gravitational field strength here is 9.8 N kg^{-1} . The trainee astronauts are in free-fall; the only force acting on them is gravity. They move with an acceleration of 9.8 m s^{-2} towards Earth and are in a state of apparent weightlessness.

Worked example 3.5B

A space shuttle is in a stable circular orbit 200 km ($2.0 \times 10^5 \text{ m}$) above the Earth's surface. An astronaut of mass 85 kg is on board. Given that the Earth has a radius of $6.4 \times 10^6 \text{ m}$ and a mass of $6.0 \times 10^{24} \text{ kg}$, calculate:

- the weight of the astronaut in this orbit
- the speed at which the shuttle orbits
- the centripetal acceleration of the astronaut
- the net force acting on the astronaut
- the apparent weight of the astronaut.

Solution

- a** The weight of the astronaut is:

$$\begin{aligned} F_g &= \frac{GMm}{R^2} \\ &= \frac{6.67 \times 10^{-11} \times 6.0 \times 10^{24} \times 85}{(6.6 \times 10^6)^2} \\ &= 782 \text{ N} \end{aligned}$$

- b** $F_g = \frac{GMm}{R^2} = \frac{mv^2}{R}$,

$$\begin{aligned} \text{so } v &= \sqrt{\frac{GM}{R}} \\ &= \sqrt{\frac{6.67 \times 10^{-11} \times 6.0 \times 10^{24}}{6.6 \times 10^6}} \\ &= 7.8 \times 10^3 \text{ m s}^{-1} \end{aligned}$$

- c** The centripetal acceleration of the shuttle is:

$$\begin{aligned} a &= \frac{v^2}{R} = \frac{(7.8 \times 10^3)^2}{6.6 \times 10^6} \\ &= 9.2 \text{ m s}^{-2} \text{ towards the centre of Earth} \end{aligned}$$

- d** The net force on the astronaut is:

$$\Sigma F = ma = 85 \times 9.2 = 780 \text{ N towards the centre of Earth}$$

- e** Since the weight force of 780 N totally accounts for the net force acting on the astronaut, the normal force must be equal to zero. In other words, the apparent weight of the astronaut is zero.



3.5 summary

Apparent weight and weightlessness

- Mass is a measure of resistance of a body to acceleration. The mass of a body is not affected by changes in the gravitational field.
- Weight, F_g or W , is the gravitational force that acts on an object and is measured in newtons. The weight of an object changes as the gravitational field strength changes.
- True weightlessness occurs when the gravitational field strength is negligible. This is possible in deep space far away from the gravitational attraction of stars and planets.
- The apparent weight of a person is equal in magnitude to the normal force, F_N or N , that the supporting surface exerts on them.
- The apparent weight of an object changes if it moves with some vertical acceleration.
- A person will be in a state of apparent weightlessness when in free-fall and moving with an acceleration equal to the gravitational field strength at their location. The person will experience zero normal force at this time.



3.5 questions

Apparent weight and weightlessness

In the following questions, you may assume that:

mass of Earth = 6.0×10^{24} kg

radius of Earth = 6.4×10^6 m

gravitational field strength on Earth's surface

= 9.80 N kg^{-1}

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

The following information applies to questions 1–3.

Gracie, of mass 25 kg, is playing on her trampoline. She has a toy of mass 0.15 kg in her hand as she jumps up and down.

- What is the weight of the toy?
 - What is the mass of the toy?
- Gracie places the toy on the palm of her outstretched hand as she performs a jump. What is the apparent weight of the toy during the jump?
- Gracie again places the toy on the palm of her outstretched hand as she performs a jump, but this time she removes her hand from under the toy during the jump. How does this affect the motion of the toy?
- A girl of mass 50 kg is riding on a 'Big Drop' thrill ride that is descending at 9.8 m s^{-2} .
 - Calculate her weight while she is descending.
 - Calculate the magnitude of the normal force acting on her.
 - What is her apparent weight?
- Question 4 is an example of how a person can experience the feeling of weightlessness without leaving the Earth.
 - Why is this often described as 'apparent weightlessness'?

- If a person is to experience weightlessness on Earth, what can you say about their acceleration compared with the acceleration due to gravity?

- During lift-off, an astronaut of mass 90 kg experiences an acceleration of 35 m s^{-2} upward. Assuming g is 9.8 N kg^{-1} , calculate:

- the weight of the astronaut
- the net force acting on the astronaut (in kN)
- the apparent weight of the astronaut (in kN).

The following information applies to questions 7–10.

A spacecraft is in a stable orbit around the Earth. The radius of this orbit is 6.8×10^6 m. An astronaut of mass 60 kg is on board the spacecraft.

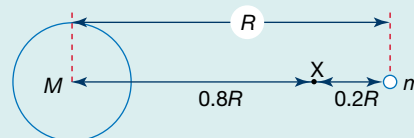
- Calculate the gravitational field strength at this radius (in N kg^{-1}).
 - Determine the weight of the astronaut in this orbit.
- Calculate the orbital speed of the spacecraft (in km s^{-1}).
 - Determine the acceleration of the astronaut while travelling in the spacecraft.
 - Calculate the net force acting on the astronaut.
- People on Earth viewing a television picture beamed from the spacecraft observed the astronaut drifting freely around the interior of the spacecraft. A typical comment from an uninformed observer might be, 'There is no gravity up there, so the astronaut is weightless'. Briefly discuss the validity of this comment and suggest an alternative comment that sums up the situation more accurately.
- Briefly explain why this astronaut experiences a feeling of weightlessness while in this stable orbit.

chapter review

In the following questions, assume that the universal constant of gravitation, $G = 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$; gravitational field strength on surface of the Earth, $g = 9.80 \text{ N kg}^{-1}$.

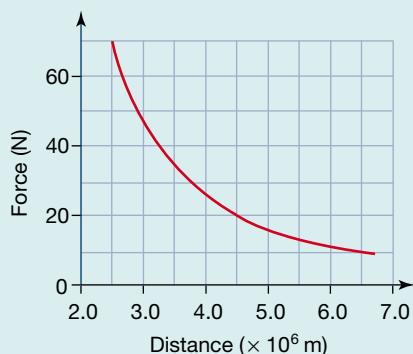
- The gravitational force of attraction between Saturn and Dione, a satellite of Saturn, is equal to $2.79 \times 10^{20} \text{ N}$. Calculate the orbital radius of Dione. Data: mass of Dione = $1.05 \times 10^{21} \text{ kg}$, mass of Saturn = $5.69 \times 10^{26} \text{ kg}$.
- A person standing on the surface of the Earth experiences a gravitational force of 900 N. What gravitational force will this person experience at a height of 2 Earth radii above the Earth's surface?
 - 900 N
 - 450 N
 - zero
 - 100 N
- During a space mission, an astronaut of mass 80 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 8.2 N kg^{-1} .
 - What is the apparent weight of the astronaut during lift-off?
 - zero
 - 780 N
 - 2400 N
 - 3200 N
 - During the lift-off phase, the astronaut will feel:
 - lighter than usual
 - heavier than usual
 - the same as usual.
 - The weight of the astronaut during the lift-off phase is:
 - lower than usual
 - greater than usual
 - the same as usual.
 - During the orbit phase, the apparent weight of the astronaut is:
 - zero
 - 780 N
 - 2400 N
 - 660 N
 - During the orbit phase, the weight of the astronaut is:
 - zero
 - 780 N
 - 2400 N
 - 660 N
- 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 .



- Neptune has a planetary radius of $2.48 \times 10^7 \text{ m}$ and a mass of $1.02 \times 10^{26} \text{ kg}$.
 - Calculate the gravitational field strength on the surface of Neptune.
 - 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.
 - 9.8 m s^{-2}
 - zero
 - 11 m s^{-2}
 - 1.6 m s^{-2}
- Given that the mass of the Earth is $5.98 \times 10^{24} \text{ kg}$ and the mean distance from the Earth to the Moon is $3.84 \times 10^8 \text{ m}$, calculate the orbital period of the Moon. Express your answer in days.
- One of Jupiter's moons, Leda, has an orbital radius of $1.10 \times 10^{10} \text{ m}$. The mass of Jupiter is equal to $1.90 \times 10^{27} \text{ kg}$. Calculate:
 - the orbital speed of Leda
 - the orbital acceleration of Leda
 - the orbital period of Leda (in days).
- Which of the following best explains what is meant by a satellite being in a geosynchronous orbit?
 - It is orbiting the Earth.
 - It is orbiting the Moon and remains above the same location.
 - It is orbiting the Earth and remains above the same location.
 - It is orbiting the Earth and returns to the same location every 24 hours.
- What is the purpose of such an orbit?
- 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 = $6.0 \times 10^{24} \text{ kg}$).
- The planet Mercury has a mass of $3.30 \times 10^{23} \text{ kg}$. Its period of rotation about its axis is equal to $5.07 \times 10^6 \text{ s}$. For a satellite to be in a synchronous orbit around Mercury, calculate:
 - the orbital radius of the satellite
 - its orbital speed
 - 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.

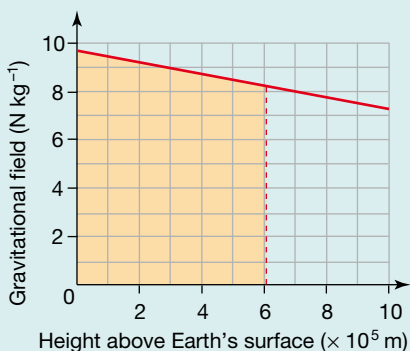


A 20 kg rock is speeding towards Mercury. When the rock is 3.0×10^6 m from the centre of the planet, its speed is estimated at 1.0 km s^{-1} . Using the graph, estimate:

- the increase in kinetic energy of the rock as it moves to a point that is just 2.5×10^6 m from the centre of Mercury
 - the kinetic energy of the rock at this closer point
 - the speed of the rock at this point
 - the gravitational field strength at 2.5×10^6 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.
- orbital period of S_1 /orbital period of S_2
 - orbital speed of S_1 /orbital speed of S_2
 - acceleration of S_1 /acceleration of S_2 .
- 12 Earth is in orbit around the Sun. The Earth has an orbital radius of 1.5×10^{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 13–17.

The diagram shows the gravitational field and distance near the Earth. A wayward satellite of mass 1000 kg is drifting towards the Earth.



- 13 What is the gravitational field strength at an altitude of 300 km?
- 14 Which of the following units is associated with the area under this graph?
- J
 - m s^{-2}
 - J s
 - J kg^{-1}
- 15 Which one of the following quantities is represented by the shaded area on the graph? (Ignore air resistance.)
- The kinetic energy per kilogram of the satellite at an altitude of 600 km
 - The loss in gravitational potential energy of the satellite
 - The loss in gravitational potential energy per kilogram of the satellite as it falls to the Earth's surface
 - The increase in gravitational potential energy of the satellite as it falls to the Earth's surface
- 16 How much kinetic energy does the satellite gain as it travels from an altitude of 600 km to 200 km altitude?
- 17 In reality, would the satellite gain the amount of kinetic energy that you have calculated in Question 16? Explain.

The following information relates to questions 18–20.

Charon is the largest of Pluto's moons. It orbits Pluto with a period of 6.4 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.

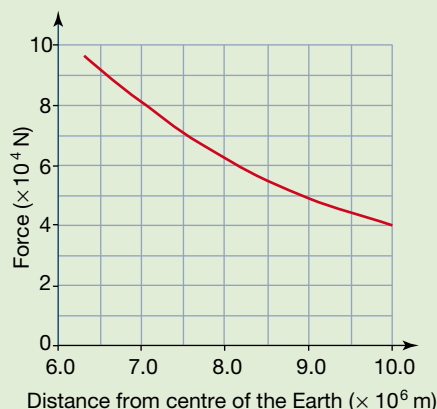
- 18 Which of the following statements is correct?
- The gravitational force between Nix and Pluto is greater than the force between Charon and Pluto.
 - The gravitational force between Nix and Pluto is less than the force between Charon and Pluto.
 - The gravitational force between Nix and Pluto is equal to the force between Charon and Pluto.
 - The gravitational forces cannot be compared with the information given.
- 19 Use Kepler's third law to determine the orbital period of Nix (in days).
- 20 Use the data relating to Charon to calculate the mass of Pluto.

exam-style questions Motion in one and two dimensions

In the following questions, assume that the acceleration due to gravity is $g = 9.8 \text{ m s}^{-2}$ and ignore the effects of air resistance.

The following information applies to questions 1–4.

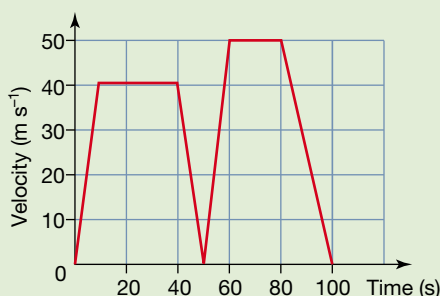
A 10 000 kg spacecraft is drifting directly towards the Earth. When it is at an altitude of 600 km, its speed is 1.5 km s^{-1} . The radius of the Earth is 6400 km. The following graph shows the force on the spacecraft against distance from the Earth.



- How much gravitational potential energy would the spacecraft lose as it falls to an altitude of 100 km?
- Determine the speed of the spacecraft at 100 km altitude.
- What is the weight of the spacecraft when it is at an altitude of:
 - 3600 km?
 - $6.0 \times 10^5 \text{ m}$?
- How does the acceleration of the spacecraft change as it moves from an altitude of 600 km to an altitude of 100 km? Include numerical data in your answer.

The following information applies to questions 4–8.

A prototype of a new model sports car of mass 1000 kg was tested on a straight length of road. An idealised version of the resulting velocity–time graph is shown in the following diagram.



- Calculate the total distance (in km) the car travelled during the 100 s interval.
- During which interval is the net force acting on the car the greatest?

- 0–10 s
- 20–30 s
- 50–60 s
- 60–70 s

b Justify your answer to part a.

- For what time interval(s) will the net force acting on this car be zero?
- At what time or times, if any, was the car travelling back towards its starting point?

The following information applies to questions 9–13.

In the Gravitron ride, the patrons enter a cylindrical chamber which rotates rapidly, causing them to be pinned to the vertical walls. A particular Gravitron ride has a radius of 5.00 m and rotates with a period of 2.50 s. Jodie of mass 60.0 kg is on the ride.

- Select the correct responses in the following statement: As the Gravitron spins at a uniform rate and the girl is pinned to the wall, the horizontal forces acting on Jodie are balanced/unbalanced and the vertical forces are balanced/unbalanced.
- Calculate the speed of Jodie as she revolves on the ride.
- What is the magnitude of her centripetal acceleration?
- Calculate the magnitude of the normal force that acts on Jodie from the wall of the Gravitron.
- The rate of rotation of the ride is increased so that Jodie completes six revolutions every 10.0 s. What is the frequency of Jodie's motion now?

The following information applies to questions 14–16.

Neptune was discovered in 1846. It has a mass of $1.03 \times 10^{26} \text{ kg}$ and a diameter of 49 500 km. Thalassa is one of the moons of Neptune. It has an orbital period of 0.31 days.

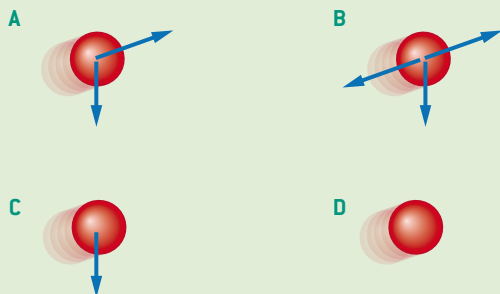
- What is the orbital radius of Thalassa?
- Calculate the speed of Thalassa (in km s^{-1}).
- Another of Neptune's moons, Proteus, has an orbital period of 1.1 days. Use Kepler's third law to determine the orbital radius of Proteus.

The following information applies to questions 17–20.

Two friends, Elvis and Kurt, are having a game of catch. Elvis throws a baseball to Kurt, who is standing 8.0 m away. Kurt catches the ball at the same height of 2.0 s after it is thrown. The mass of the baseball is 250 g.

- Determine the value of the maximum height gained by the ball during its flight.
- What was the acceleration of the ball at its maximum height?
- Calculate the speed at which the ball was thrown.

- 20 Which of the following diagrams best shows the forces acting on the ball just after it has left Elvis's hand?



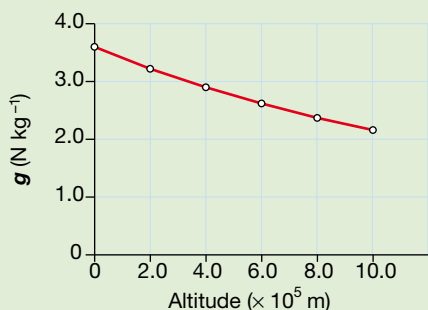
The following information applies to questions 21–24.

A cricket ball of mass 200 g moves towards a batsman at a velocity of 20 m s^{-1} north. After the ball is struck by the batsman, the velocity is 30 m s^{-1} south. The interaction between bat and ball lasts just 10 ms.

- 21 Determine the change in momentum of the ball during this time.
- 22 Calculate the net impulse that the bat delivered to this ball.
- 23 a Determine the average force that the bat exerted on the ball during the interaction.
- b Select the correct response in this statement: The force that the ball exerts on the bat is smaller in magnitude than/equal to/greater in magnitude than the force that the bat exerts on the ball.
- 24 If the net force–time graph were plotted for the interaction between the bat and the ball, which one or more of the following units would apply to the area under the graph?
- A J B m s^{-2} C kg m s^{-1}
 D kg m s^{-2} E N s

The following information applies to questions 25–27.

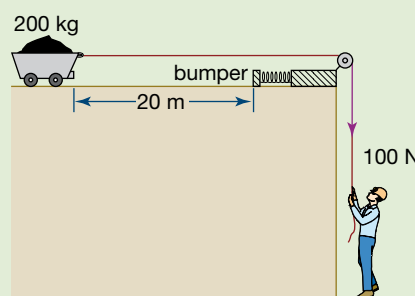
A small asteroid has just smashed into the surface of Mars and a lump of Martian rock of mass 20 kg has been thrown into space with 40 MJ of kinetic energy. A graph of gravitational field–distance from the surface of Mars is shown below.



- 25 What is the gravitational force acting on the Martian rock when it is at an altitude of 300 km?
- 26 How much kinetic energy (in MJ) does the rock lose as it travels from the surface of Mars to an altitude of $6.0 \times 10^5 \text{ m}$?
- 27 The rock eventually comes to a stop and starts to fall back towards Mars. Without actually doing the calculations, explain how you would determine the altitude at which the rock stopped.

The following information applies to questions 28–34.

A small-time gold prospector sets up a cable-pulley system that allows him to move a container full of ore of total mass 200 kg from rest a distance of 20 m along a level section of rail track as shown in the following diagram.



When the load reaches the end of the track, it is momentarily brought to rest by a powerful spring-bumper system, which is assumed to have negligible mass. A constant frictional force of 30 N acts on the wheels of the container-car along the track. Assume that there is negligible friction between the pulley and the cable. The prospector applies a constant force of 100 N to the rope as the trolley moves along the track.

- 28 How much work is done by the prospector in moving the 200 kg load along the track?
- 29 How much work is done on the container as it moves along the track?
- 30 Calculate the change in kinetic energy of the load as it moves along the track.
- 31 What is the speed of the load when it reaches the end of the track?
- 32 How much energy is converted into heat as the load moves over the track?
- 33 What is the power output of the prospector as he moves one 200 kg load over the track during a 10 s interval?
- 34 The spring bumper has a force constant of 150 N m^{-1} . How much kinetic energy does the container lose as the spring compresses by 7.5 cm?

The following information applies to questions 35–37.

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. Data: radius of Earth = 6.4×10^6 m, mass of Earth = 6.0×10^{24} kg.

- 35 Determine the value of the ratio: speed of ISS/speed of Optus D2.
 36 Determine the value of the ratio: period of ISS/period of Optus D2.
 37 Determine the value of the ratio: acceleration of ISS/acceleration of Optus D2.

The following information applies to questions 38–42.

A physics student decides to study the properties of a bungee rope by recording the extension produced by various masses attached to the end of a section of the rope. The results of the experiment are shown in the following table.

Mass (kg)	Extension (m)
0.5	0.24
1.0	0.52
1.5	0.73
2.0	0.95
2.5	1.20
3.0	1.48
3.5	1.70

- 38 Draw the force versus extension graph for this bungee rope.
 39 Estimate the value of the force constant for this rope.
 During an investigation, the student stretched the rope horizontally by 15 m.
 40 Assuming that the rope behaves ideally, determine the potential energy stored in the bungee rope at this point.
 Finally, the student stands on a skateboard and allows the stretched rope to drag her across the smooth floor of the school gymnasium.
 41 Which statement best describes the motion of the student?
 A She moves with a constant velocity.
 B She moves with a constant acceleration.
 C She moves with increasing velocity and decreasing acceleration.
 D She moves with increasing velocity and increasing acceleration.
 42 The combined mass of the student and her board is 60 kg. Calculate the maximum speed that she attains as she is pulled by the bungee cord.

The following information applies to questions 43–45.

At football training, some of the players are throwing themselves at a large tackle bag of mass 45 kg. During one exercise, a ruckman of mass 120 kg running at 6.0 m s^{-1} crashes into the stationary bag and carries it forward.

- 43 What is the combined speed of the bag and ruckman?
 44 How much momentum does the ruckman lose?
 45 How much momentum does the tackle bag gain?
 46 Select the correct responses in these statements: The collision between the ruckman and tackle bag is elastic/inelastic. Momentum/kinetic energy has been conserved and momentum/kinetic energy has been transformed into heat and sound energy.

The following information applies to questions 47–48.

A skateboarder of mass 55 kg is practising on a half-pipe of radius 2.0 m. At the lowest point of the half-pipe, the speed of the skater is 6.0 m s^{-1} .

- 47 a What is the acceleration of the skater at this point?
 b Calculate the size of the normal force acting on the skater at this point.
 48 Discuss the apparent weight of the skater as they travel through the lowest point in the pipe.

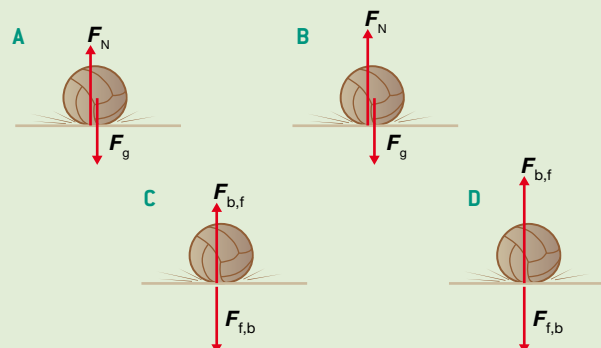
The following information applies to questions 49–50.

A netball is dropped vertically from a height of 1.5 m onto a horizontal floor. The diagrams below relate to the instant that the ball reaches the floor and is stationary for a short period of time before rebounding.

- 49 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.



- 50 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.

Unit 2

area of study 2

Electronics and photonics

outcome

On completion of this area of study, you should be able to investigate, describe, compare and explain the operation of electronic and photonic devices, and analyse their use in domestic and industrial systems.

Electronics

Electronics is one of the main drivers of modern technology. In a very real sense, many of the technological advances of the 20th century were made possible because of rapid advancements in electronics. Continual miniaturisation of electronic components (especially complex transistor circuits) has meant that the cost of many electronic devices has decreased while their functionality has dramatically increased. Computers are a perfect example—they are very complex electronic devices. New computers get more and more powerful and include more and more features, while staying the same price (in relative terms) or becoming cheaper. Continual miniaturisation of electronic components has made this possible.

Many of our common household appliances, such as fridges, washing machines and air conditioners, are controlled by microcontrollers—small but powerful single-chip computers.

We shall revise some key concepts in electronics that you will probably be familiar with, and introduce some new concepts and devices. We shall then use these concepts and devices to analyse and interpret the operation of some simple electronic circuits.

by the end of this chapter

you will have covered material from the study of electronics, including:

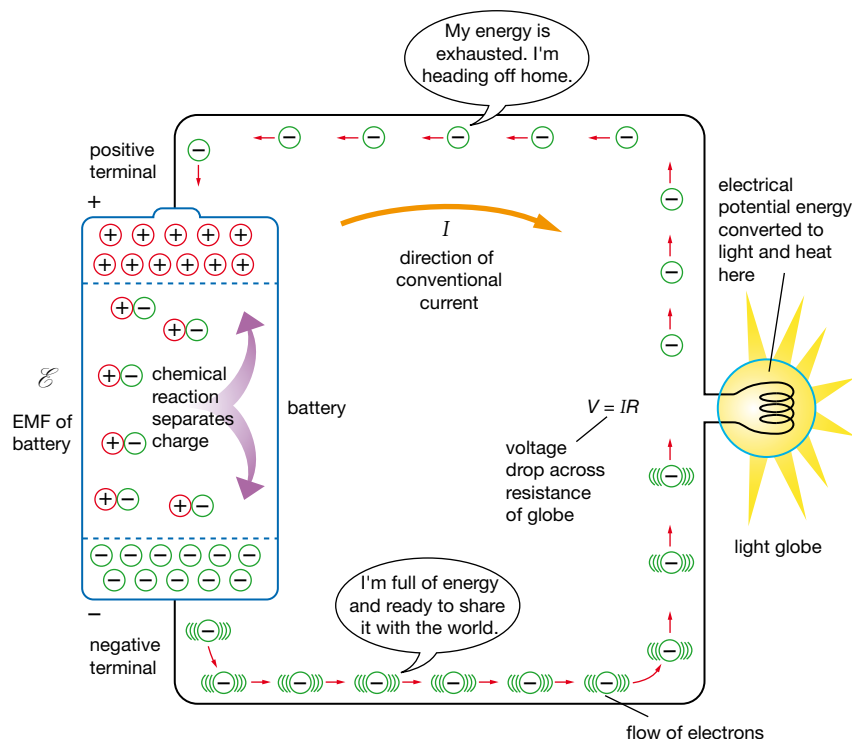
- concepts of current, resistance, potential difference (voltage drop) and power as applied to the operation of simple DC electronic circuits
- the operation of diodes, resistors and thermistors
- Ohm's law, $V = IR$
- power, $P = IV$.

4.1 Analysing electronic circuits

Electronics review

In Year 11 you studied how the chemical reactions inside a battery separate positive and negative electric charges, converting energy released by chemical reactions to electrical potential energy between the battery terminals. This potential energy can be converted to other useful forms of energy (heat, light, sound etc.) when an electronic circuit is connected across the terminals. The electronic circuit provides a conducting path between the oppositely charged battery terminals, and mobile charge can therefore move from one terminal to the other under the influence of the electric force (giving up its electrical potential energy in the process). It is important to realise that the electrical potential energy is the only thing that decreases as the charge carriers move around the circuit—the number of carriers and the charge on each carrier always remain constant (i.e. charge carriers are *not* used up). Remember that instead of referring to the total electrical potential energy between two points in a circuit, in electronics it is more convenient to talk about the electric potential (potential energy per unit charge). This is often simply called potential difference or *voltage*. For a battery, the *EMF* is the energy supplied per unit charge by the chemical reactions responsible for the charge separation. The SI unit for both electric potential and EMF is the volt (V).

It is convenient to choose a *zero electric potential reference point* when analysing electronic circuits. This reference point is often called the *circuit ground* or *earth point*. Any voltage that appears to refer to only one specific point in a circuit (rather than the difference between two points) can be assumed to be measured with respect to circuit ground.



The charge carriers that transport the electrical energy throughout the circuit are usually electrons, although in certain circumstances they can be positive ions or a combination of both. For this reason, when physically



PRACTICAL ACTIVITY 17

Internal resistance of a battery

Physics file

An ideal battery maintains a constant **EMF** regardless of how much current is drawn from it. In practice, the potential difference across the terminals of a real battery decreases as the current drawn from the battery increases. The decrease in terminal voltage becomes significant when relatively large currents are drawn from the battery. The phenomenon is attributed to the contact resistance, ionic resistance of the chemical cell and the limitation of the rate at which charge can be supplied by the chemical reaction inside the battery. In electronic circuits, we often combine all these limitations (of a real battery) and model them as one internal resistance in series with an ideal battery, as shown in Figure 4.2.

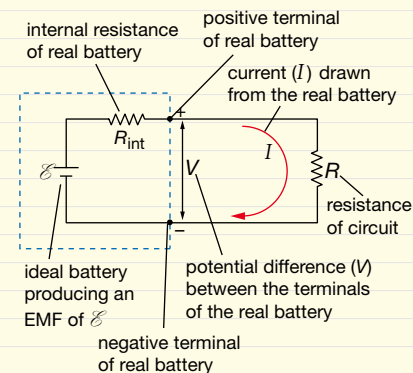


Figure 4.2 A battery with internal resistance.

Figure 4.1 A battery connected to a light globe. The large arrow shows the direction of 'conventional' current; the movement of electrons is in the opposite direction. Electrons have a lot of potential energy as they leave the negative battery terminal but little as they return to the positive terminal. Energy is converted to light and heat in the light globe. The current leaving the battery equals the current returning to the battery.

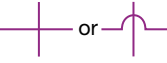

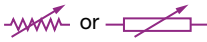


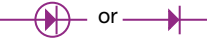

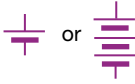
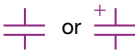






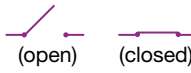

Wires crossed not joined	
Wires crossed at a junction	
Variable resistor	
Fixed resistor	
Unspecified circuit element	
Diode	
Earth or ground	
Battery or DC supply	
Capacitor	
Time-varying or AC supply	
Ammeter	
Voltmeter	
Ohmmeter	
Transistor (npn)	
Thermistor or temperature-dependent resistor	
Switch	
Lamp/bulb	

Figure 4.3 Summary of circuit symbols used in Chapter 4.

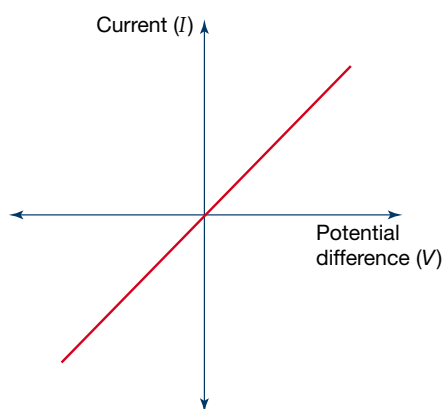


Figure 4.4 Current versus potential difference for an ohmic resistor.

explaining how electronic devices work, it is often convenient to talk about the *flow* of the actual positive or negative charge around a conducting path. In electronic circuits, we talk about conventional electric current (I), which is defined in terms of the transfer of positive charge with respect to time. Hence:

$$I = \frac{Q}{t}$$

The direction of conventional current around a conducting path is therefore *defined* to be the same as the direction that positive charge would flow around the circuit. If the charge carriers happen to be positive, then the directions of mobile charge flow and conventional current are the same. If the charge carriers are negative (i.e. electrons), then the direction of electron flow is opposite to that of conventional current. In simple electronic circuits, the direction of conventional current goes from the positive to negative battery terminals, although the actual flow of electrons is in the opposite direction (see Figure 4.1).

The circuit symbols for the electronic devices introduced in this chapter are summarised in Figure 4.3.

The energy stored or dissipated in any electronic device is equal to the potential difference across it multiplied by the charge flowing through it ($U = VQ$). Hence, it follows that the power dissipated in any device (power is the rate of change of energy with respect to time) is equal to the potential difference multiplied by the current:

$$P = \frac{U}{t} = V\left(\frac{Q}{t}\right) = VI$$

The conductors and resistors you studied in Year 11 were mainly ohmic resistors; that is, they behaved according to Ohm's law. The graph of current versus voltage for any ohmic device is shown in Figure 4.4. The resistance of a circuit element determines the amount of current able to flow when a given potential difference is applied. An ohmic device has a *constant* resistance, regardless of the applied voltage. Most resistors and metal conductors at a constant temperature are ohmic devices.



OHM'S LAW states that the current flowing in a conductor is directly proportional to the potential difference across it, i.e. $I \propto V$. The constant of proportionality is called the resistance:

$$V = IR$$

Using Ohm's law together with the expression of power, we can show that:

$$\begin{aligned} P &= VI \\ &= I^2 R \\ &= \frac{V^2}{R} \end{aligned}$$

In Year 11 you also learned that the equivalent resistance of any number of resistors in series (see Figure 4.5a) is given by:

$$R_{\text{eq}} = R_1 + R_2 + R_3 + \dots + R_n$$

and in a series circuit:

$$\mathcal{E} \text{ or } V_{\text{supply}} = V_1 + V_2 + V_3 + \dots$$

and

$$I_{\text{total}} = I_1 = I_2 = I_3 = \dots$$

Also recall that the equivalent resistance of any number of resistors in parallel (see Figure 4.5b) is given by:

$$\frac{1}{R_{\text{eq}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \dots + \frac{1}{R_n}$$

and in a parallel circuit:

$$\mathcal{E} \text{ or } V_{\text{supply}} = V_1 = V_2 = V_3 = \dots$$

and

$$I_{\text{total}} = I_1 + I_2 + I_3 + \dots$$

Worked example 4.1A

Consider the circuit shown in the diagram.

- a** If $\mathcal{E} = 12 \text{ V}$, $R_1 = 10 \text{ k}\Omega$, $R_2 = 5 \text{ k}\Omega$ and $R_3 = 8 \text{ k}\Omega$, calculate the voltage across R_2 .
- b** If $\mathcal{E} = 10 \text{ V}$, $V_1 = 3 \text{ V}$, $R_2 = 10 \text{ k}\Omega$ and $R_3 = 4 \text{ k}\Omega$, calculate the resistance of R_1 .

Solution

- a** The total resistance of the series circuit is:

$$\begin{aligned} R_t &= R_1 + R_2 + R_3 \\ &= 10 + 5 + 8 \\ &= 23 \text{ k}\Omega \end{aligned}$$

and so the current flowing around the conduction path is:

$$\begin{aligned} I &= \frac{\mathcal{E}}{R_t} \\ &= \frac{12}{23 \times 10^3} \\ &= 5.2 \times 10^{-4} \\ &= 0.52 \text{ mA} \end{aligned}$$

and hence the voltage across R_2 is:

$$\begin{aligned} V_2 &= IR_2 \\ &= (0.52 \times 10^{-3})(5 \times 10^3) \\ &= 2.6 \text{ V} \end{aligned}$$

- b** The total resistance of R_2 in series with R_3 is:

$$\begin{aligned} R_t &= R_2 + R_3 \\ &= 10 + 4 \\ &= 14 \text{ k}\Omega \end{aligned}$$

The total voltage across the two resistors is:

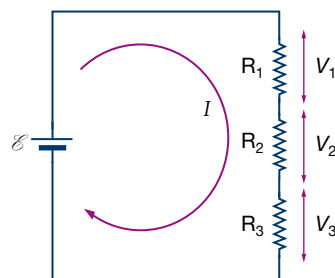
$$\begin{aligned} V_t &= \mathcal{E} - V_1 \\ &= 10 - 3 \\ &= 7 \text{ V} \end{aligned}$$

Hence the current flowing around the conduction path is:

$$\begin{aligned} I &= \frac{V_t}{R_t} \\ &= \frac{7}{14 \times 10^3} \\ &= 5.0 \times 10^{-4} \\ &= 0.5 \text{ mA} \end{aligned}$$

and the resistance of R_1 is:

$$\begin{aligned} R_1 &= \frac{V_1}{I} \\ &= \frac{3}{0.5 \times 10^{-3}} \\ &= 6 \text{ k}\Omega \end{aligned}$$



(a) I same current through all resistors connected in series

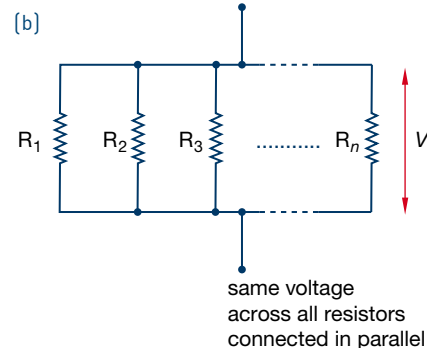
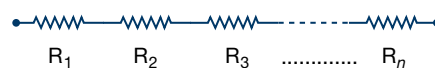


Figure 4.5 (a) Resistors in series; (b) resistors in parallel.

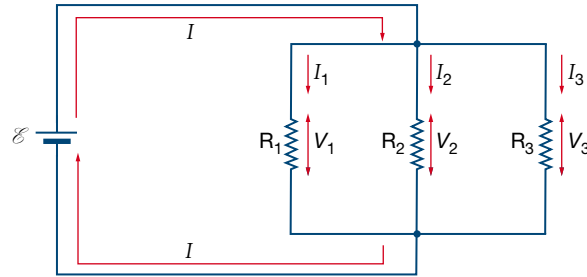
Physics file

The total current flowing from the battery terminals and into the circuit is called the *line current*.

The current flowing in any particular branch of the circuit is called a *branch current*.

Worked example 4.1B

Consider the circuit shown in the diagram.



- a** If $\mathcal{E} = 15 \text{ V}$, $R_1 = 10 \text{ k}\Omega$, $R_2 = 5 \text{ k}\Omega$ and $R_3 = 8 \text{ k}\Omega$, calculate the total current flowing through the circuit.
- b** If $\mathcal{E} = 10 \text{ V}$, $I = 10 \text{ mA}$, $R_2 = 10 \text{ k}\Omega$ and $R_3 = 4 \text{ k}\Omega$, calculate the resistance of R_1 .

Solution

- a** The total resistance of the parallel circuit is:

$$\begin{aligned}\frac{1}{R_t} &= \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \\ &= \frac{1}{10 \times 10^3} + \frac{1}{5 \times 10^3} + \frac{1}{8 \times 10^3}\end{aligned}$$

and therefore $R_t = 2.4 \text{ k}\Omega$

Hence the total current flowing around the conduction path is:

$$\begin{aligned}I &= \frac{\mathcal{E}}{R_t} \\ &= \frac{15}{2.4 \times 10^3} \\ &= 6.3 \times 10^{-4} \\ &= 0.63 \text{ mA}\end{aligned}$$

- b** The total current through R_2 and R_3 is:

$$\begin{aligned}I_2 + I_3 &= \frac{\mathcal{E}}{R_2} + \frac{\mathcal{E}}{R_3} \\ &= \frac{10}{10 \times 10^3} + \frac{10}{4 \times 10^3} \\ &= 1 \times 10^{-3} + 2.5 \times 10^{-3} \\ &= 3.5 \times 10^{-3} \\ &= 3.5 \text{ mA}\end{aligned}$$

Thus the current flowing through R_1 is:

$$\begin{aligned}I_1 &= I - (I_2 + I_3) \\ &= 10 - 3.5 \\ &= 6.5 \text{ mA}\end{aligned}$$

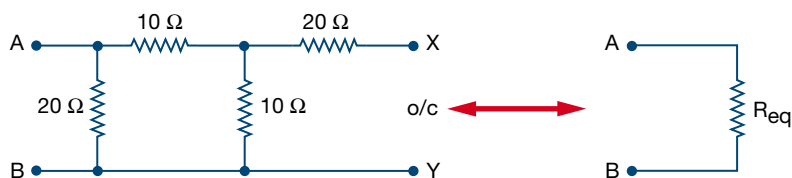
Hence the resistance of R_1 is:

$$\begin{aligned}R_1 &= \frac{\mathcal{E}}{I_1} \\ &= \frac{10}{6.5 \times 10^{-3}} \\ &= 1.5 \text{ k}\Omega\end{aligned}$$

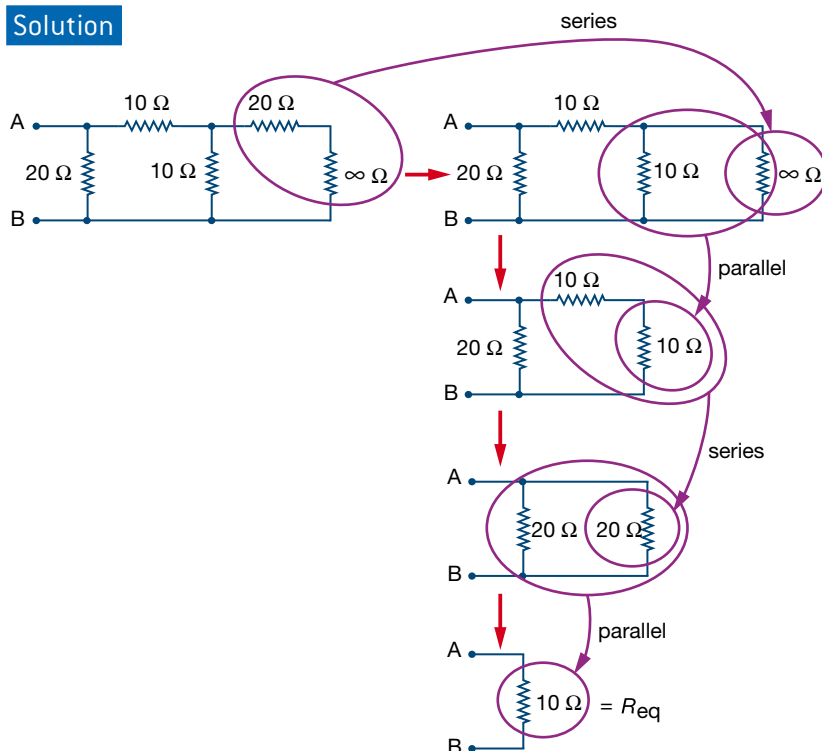
Worked example 4.1C

Simplify the following resistor network (between points A and B) to one single equivalent resistor (R_{eq}), by using the idea of series and parallel circuit simplification.

(Hint: Remember that there is no conduction path (and hence no current flow) between the points X and Y. This is often called an open circuit (o/c). As the current is zero, the resistance between the points X and Y is effectively infinite, i.e. $R_{xy} = \infty \Omega$.)



Solution



Voltage dividers

When we have a potential difference across a number of resistors in series, we can work out the voltage across a particular individual resistor by using the *voltage division principle*. For example, look at the two-resistor series circuit shown in Figure 4.6. Clearly the potential difference (V_{in}) between points A and B is the voltage that exists across the two resistors R_1 and R_2 . In this circuit, the same current (I) flows through both resistors. Hence, using Ohm's law, we see that the voltage drop across each resistor is directly proportional to its resistance, and that the voltage drop across both resistors is proportional to the total resistance. If one of the resistors has a larger resistance than the other, there will be a greater fraction of V_{in} across the larger resistor.

In mathematical terms, using Ohm's law, we have:

$$V_{in} = I(R_1 + R_2) \text{ and } V_{out} = IR_2$$

Eliminating I gives:

$$\frac{V_{out}}{V_{in}} = \frac{R_2}{R_1 + R_2}$$

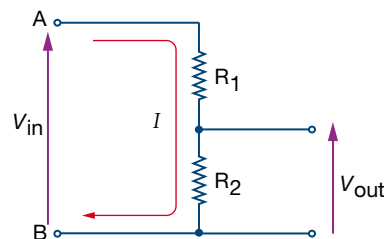


Figure 4.6 Simple two-resistor voltage divider.

Physics file

The ratio of voltage output to voltage input of a circuit is called its *voltage gain*. The relationship between output and input voltages is studied again when we look at amplifiers.

As we are often interested in the output voltage of a voltage-dividing circuit, this relationship is usually written as:



$$V_{\text{out}} = \frac{R_2}{R_1 + R_2} \times V_{\text{in}}$$

where V_{out} = output potential (V)

V_{in} = input potential (V)

R_2 = resistance across which the output potential is measured (Ω)

Note: This rule must be adjusted if V_{out} is taken from R_1 rather than R_2 .

The voltage division principle can be used for series circuits with more than *two* resistors, and in general we have:

$$V_x = \frac{R_x}{R_1 + R_2 + \dots + R_n} \times V_{\text{in}}$$

where V_x is the voltage across resistor R_x .

Worked example 4.1D

Find the voltage across R_2 in the three-resistor circuit shown.

Assume $R_1 = 2 \text{ k}\Omega$, $R_2 = 3 \text{ k}\Omega$, $R_3 = 5 \text{ k}\Omega$ and $V_{\text{in}} = 10 \text{ V}$.

Solution

$$\begin{aligned} V_{\text{out}} &= \frac{R_2}{R_1 + R_2 + R_3} V_{\text{in}} \\ &= \frac{3 \times 10^3}{(2 + 3 + 5) \times 10^3} \times 10 \\ &= 3 \text{ V} \end{aligned}$$

Note that R_2 has 30% of the total resistance of the series circuit and therefore quite logically uses 30% of the supply voltage! Using this approach can allow you to calculate V_{out} values quite quickly in voltage dividing circuits. Mathematically, this idea can be expressed as:

$$\frac{V_2}{V_{\text{in}}} = \frac{R_2}{R_{\text{t}}}$$

for series circuits.

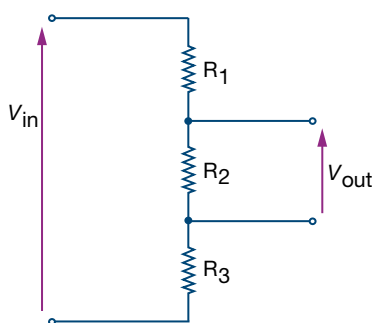
Worked example 4.1E

Determine the effective resistance for the circuit shown. Then find the total current drawn by the circuit.

Solution

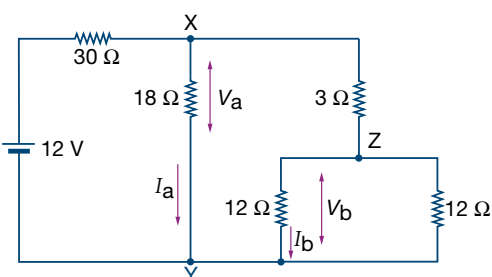
Any simple resistor network can be replaced with one single resistor that has the same overall characteristics as the original circuit, i.e. one resistor that draws exactly the same current from the battery as the original network.

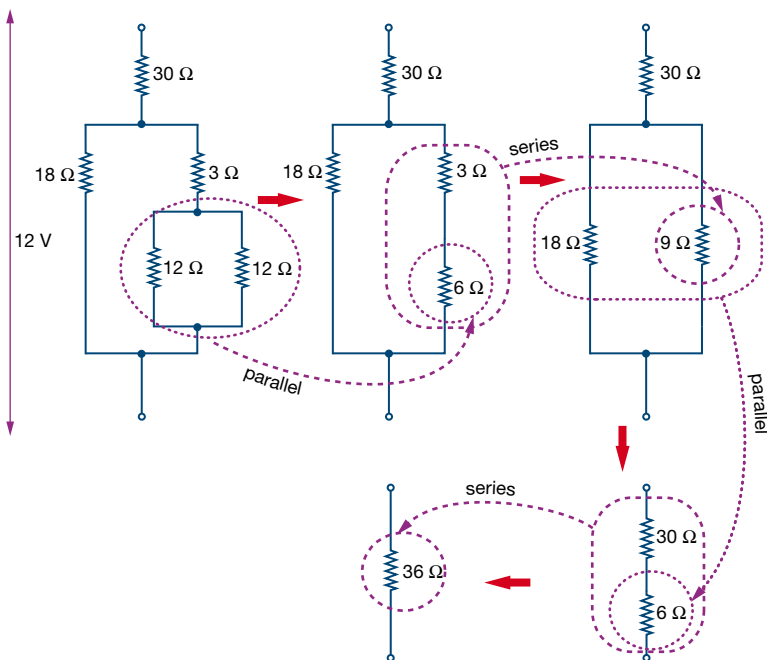
Redraw the circuit in a linear form, so that the voltage across it drops from top to bottom. This can help visualise what is going on. Now use the series and parallel resistor rules to simplify the network. Eventually you are left with one single resistor that is equivalent to the whole circuit.



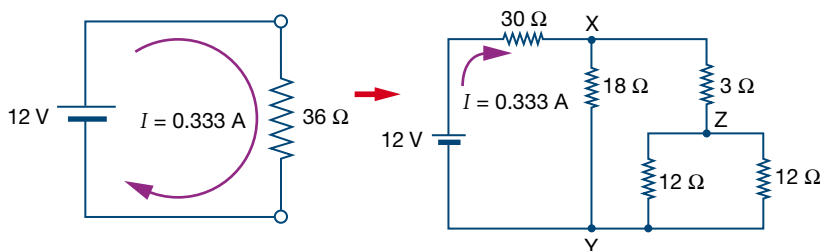
PRACTICAL ACTIVITY 18

Resistance in a combination circuit





So the entire network is equivalent to one $36\ \Omega$ resistor [i.e. this single equivalent resistor behaves exactly like the original network as far as the power supply is concerned]. Using Ohm's law, we see that the current flowing from the battery and into the network (or its equivalent resistance) is $I = 0.333\ \text{A}$.



Thermistors and voltage dividing

Voltage dividing circuits allow us to manipulate the size of a voltage in response to a given input. The principles of voltage dividing will be very important later in this chapter when we examine amplifier circuits and in the next chapter when we look at electrical-optical circuits. Voltage dividing circuits can allow us to detect and control a wide range of physical phenomena. Can you think of any electronic devices in your home that respond to light levels, air temperature, pressure or movement? It is very likely that they use the principle of voltage division to create an output voltage in response to some type of input.

A simple circuit for an electronic temperature sensor is shown in Figure 4.7. It contains a *thermistor*, which is a device whose resistance decreases sharply as its temperature rises. Its symbol is shown in Figure 4.3, page 118. The thermistor is placed in series with a selected resistor and a power supply. The potential difference of the supply is shared and the output for the circuit (V_{out}) is recorded across R_2 . When the temperature is low, the thermistor R_1 will have a high resistance and therefore use a larger share of the supply

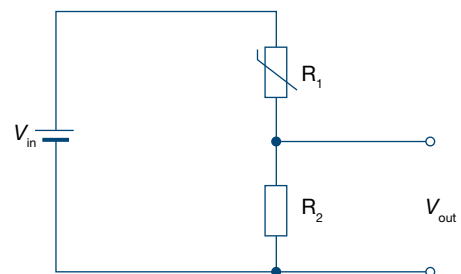


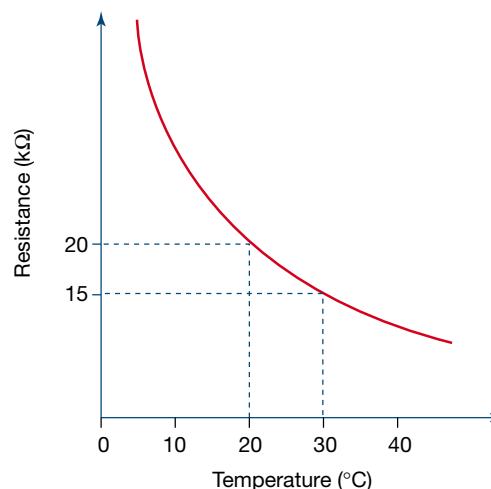
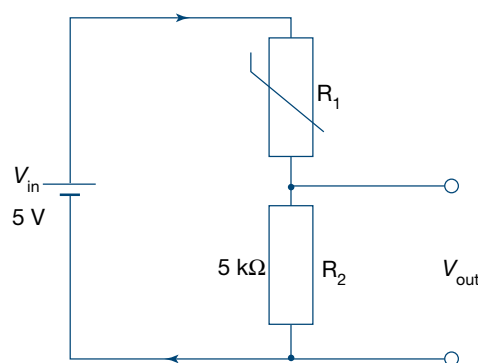
Figure 4.7 This circuit contains a thermistor.

voltage. This means that V_{out} will be low. Alternatively, if the temperature rises, the thermistor will have lower resistance and therefore use a smaller share of the supply voltage. Thus, V_{out} increases.

Worked example 4.1F

A potential divider circuit includes a thermistor and is used as a temperature sensor. The resistance–temperature characteristics of the thermistor are shown. Using this graph and the information included on the circuit diagram, determine:

- the resistance of the thermistor at 30°C
- the current in the circuit
- the output potential difference, V_{out}
- the output potential if the temperature falls to 20°C .



Physics file

The potential divider relationship assumes equal current in both resistors, i.e. no current is drawn by the voltmeter measuring the output potential. In fact, voltmeters always draw a small current, making it difficult to draw accurate predictions of resistance values. This ‘loading’ effect can amount to significant errors and should be considered during any practical investigations.

Solution

- The resistance of the thermistor can be found directly from the resistance–temperature graph. At 30°C the thermistor would have a resistance of $15\text{ k}\Omega$.
- $$I = \frac{V_{\text{in}}}{R_1 + R_2} = \frac{5}{15 \times 10^3 + 15 \times 10^3} = \frac{5}{20 \times 10^3} = 2.5 \times 10^{-4} \text{ A}$$
- $$V_{\text{out}} = \frac{R_2}{R_1 + R_2} \times V_{\text{in}}$$
$$= \frac{5 \text{ k}\Omega}{15 \text{ k}\Omega + 5 \text{ k}\Omega} \times 5 = 1.25 \text{ V}$$
- At 20°C the thermistor's resistance will be $20\text{ k}\Omega$ and
$$V_{\text{out}} = \frac{5 \text{ k}\Omega}{20 \text{ k}\Omega + 5 \text{ k}\Omega} \times 5 = 1 \text{ V}$$

Using a multimeter for accurate measurements

As well as *calculating* the various voltages and currents in a circuit, we also need to be able to measure these values accurately. In electronics we do this with measuring instruments such as the voltmeter, ammeter and ohmmeter. These devices give a numeric readout of a particular voltage, current or resistance in a DC circuit. The numeric readout can also represent RMS values of sinusoidal voltages and currents in AC circuits. These various functions can be combined into one instrument called a multimeter.

The act of measuring a voltage or current in an electronic circuit will inevitably cause some change in that circuit, since some electrical energy must be diverted from the circuit into the measuring instrument. Multimeters are

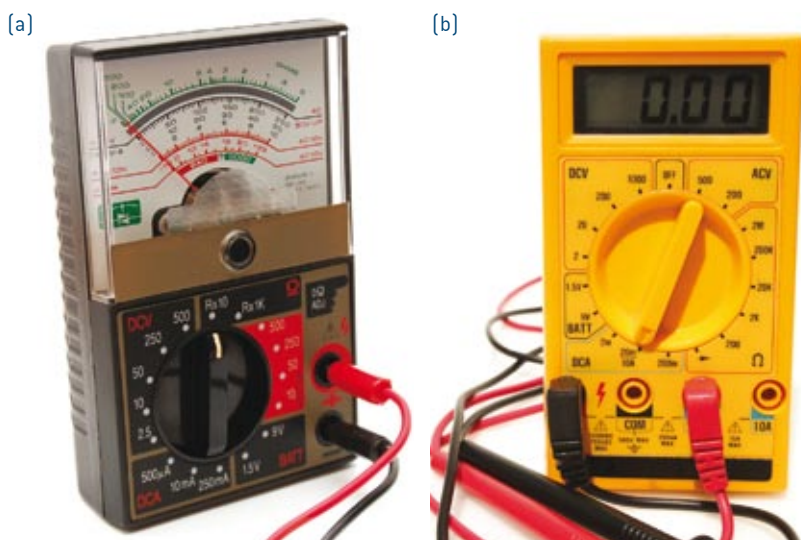


Figure 4.8 A multimeter combines the function of a voltmeter and an ammeter. Both analogue [a] and digital [b] forms are available. Although multimeters have been replaced by computer-based systems in some applications, their ease of use and portability still make them useful for making quick measurements of instantaneous circuit conditions.

very cleverly designed so as to minimise this energy diversion, but we still need to understand how multimeters work so that we can use them correctly.

Most multimeters draw a very small current from the circuit. This current is used to sense the quantity being measured and is then converted to some form of visual indicator, usually either a moving pointer (analogue display) or a digital screen (digital display) as shown in Figure 4.8.

Measuring voltage

Figure 4.9 shows how to use a multimeter to measure the *voltage* between two points in a circuit.

- The multimeter should be switched to the appropriate range for the measurement. If this is unknown, always start on the least sensitive (largest) range. Note that some multimeters are auto-ranging.
- The multimeter leads must be connected *in parallel* across the two points whose voltage difference we wish to measure.

In voltmeter mode, the multimeter has a *very large resistance between its two terminals*, so that it draws very little current from the circuit and hence has very little effect on the voltage being measured.

Measuring current

Figure 4.10 shows how to use a multimeter to measure the *current* flowing past a particular point in a circuit.

- If the multimeter is not auto-ranging, it should be switched to the appropriate range for the measurement. If this is unknown, always start on the least sensitive (largest) range.
- The multimeter leads must be connected *in series* with the electronic component through which we wish to measure the current.

In ammeter mode, the multimeter has a *very small resistance between its two terminals* so that it has very little effect on the current being measured.

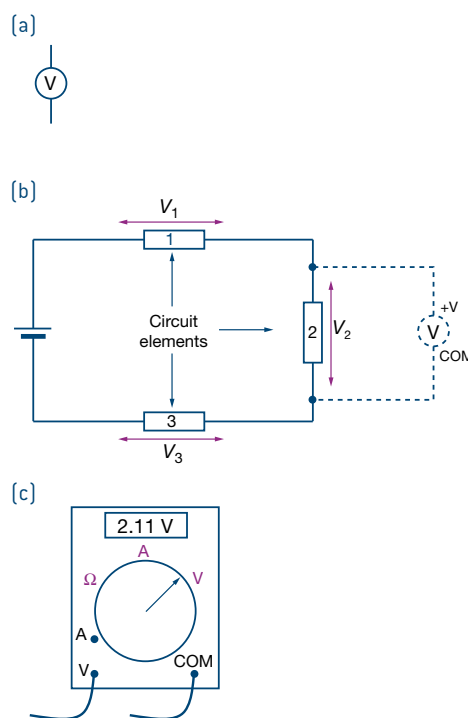


Figure 4.9 (a) Circuit symbol for a voltmeter. (b) Using a voltmeter to measure voltage between two points in a circuit. (c) A multimeter in voltmeter mode.

Physics file

An ideal voltmeter does not affect the circuit under measurement. An ideal voltmeter has an infinite resistance between its terminals.

Physics file

An *ideal* ammeter does not affect the circuit under measurement. An ideal ammeter has zero resistance between its terminals.

Physics file

Many multimeters provide a choice of scales so they can be used to measure a large range of voltage, current or resistance. For example, in voltmeter mode, a multimeter might have ranges of 0–3, 0–10 and 0–30 V etc. Since multimeters are usually used to read unknown voltage, current or resistance, it is a useful precaution to always set meters on the highest available range first. Read the value and then check if it could be measured on a lower range. Using a lower range will enable the value to be measured with greater accuracy.

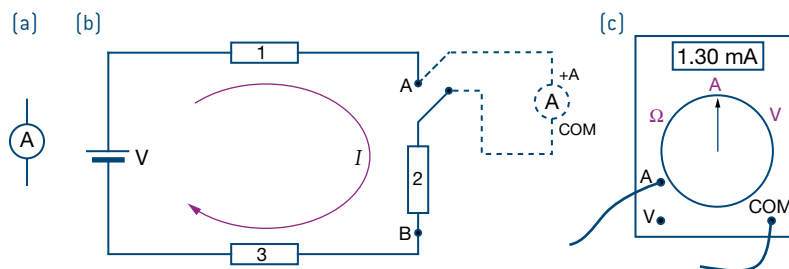


Figure 4.10 (a) Circuit symbol for an ammeter. (b) Using an ammeter to measure the current flowing through a point in a circuit. (c) A multimeter in ammeter mode.

Measuring resistance

Figure 4.11 shows how to use a multimeter to measure the *resistance* of a particular component in a circuit.

- If the multimeter is not auto-ranging, it should be switched to the appropriate range for the measurement. If this is unknown, always start on the least sensitive (largest) range.
- One end of the electronic component must be *disconnected* (and therefore isolated) from the rest of the circuit. The multimeter leads are then connected *in parallel* across this component.

The meter has a battery that passes a very small current through the unknown resistive component. Ohm's law tells us that, since the battery has a constant voltage, the current flowing through the resistive element will be inversely proportional to the unknown resistance. The meter senses the magnitude of this current and displays the corresponding resistance.

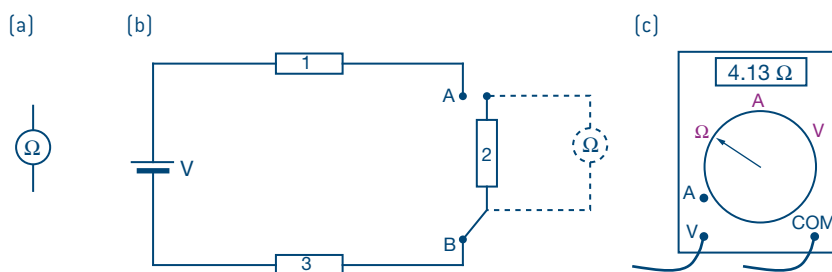


Figure 4.11 (a) Circuit symbol for an ohmmeter. (b) Using an ohmmeter to measure the resistance between two points that are disconnected from the circuit. (c) A multimeter in ohmmeter mode.



4.1 summary

Analysing electronic circuits

- The direction of conventional current is defined as the same direction that positive charge would flow in an electrical circuit.
- The operation of electronic circuits can be analysed using the following relationships:

$$I = \frac{Q}{t}$$

$$V = IR$$

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

- For a series circuit:

$$R_{\text{eq}} = R_1 + R_2 + R_3 + \dots + R_n$$

$$V_{\text{supply}} = V_1 + V_2 + V_3 + \dots$$

$$I_t = I_1 = I_2 = I_3 = \dots$$

- For a parallel circuit:

$$\frac{1}{R_{\text{eq}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \dots + \frac{1}{R_n}$$

$$V_{\text{supply}} = V_1 = V_2 = V_3 = \dots$$

$$I_t = I_1 + I_2 + I_3 + \dots$$

- The voltage division principle states:

$$V_x = \frac{R_x}{R_1 + R_2 + \dots + R_n} \times V_{\text{in}}$$

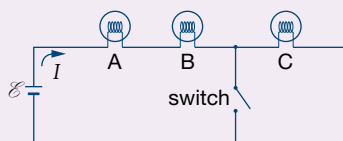


4.1 questions

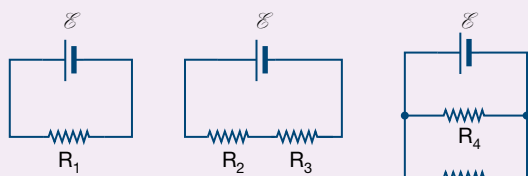
Analysing electronic circuits

- 1 A circuit consists of three identical lamps connected to a battery with a switch, as shown in the diagram. When the switch is closed, what happens to the:

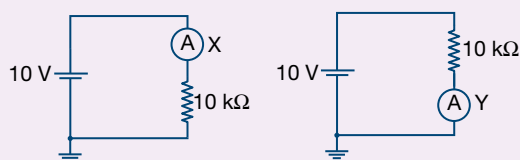
- intensity of lamp A?
- intensity of lamp C?
- current in the circuit?
- potential difference across lamp B?
- potential difference across lamp C?
- total power dissipated in the circuit?



- 2 In the following three circuits the batteries and resistors are all identical. Assume the batteries are ideal (i.e. no internal resistance).



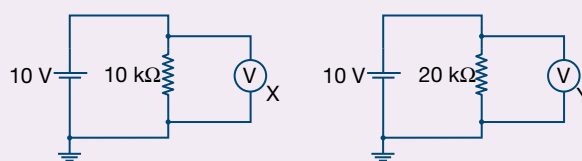
- Which resistor(s) has the highest current flowing through it?
 - Which resistor(s) has the lowest current flowing through it?
 - Which resistor(s) has the highest power dissipated through it?
- 3 a For the following circuits, which of the following statements is correct? Assume that the ammeters are identical.



- Ammeter X displays the highest current.
- Ammeter Y displays the highest current.
- Both ammeters display the same current.

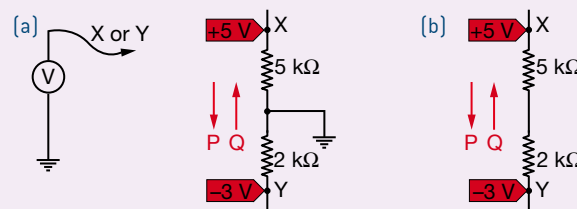
- Thermistors are devices whose resistance decreases sharply with increases in temperature.

- b For the following circuits, which of the following statements is correct? Assume that the voltmeters and batteries are ideal and identical.

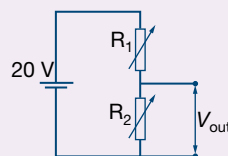


- Voltmeter X displays the higher voltage drop.
- Voltmeter Y displays the higher voltage drop.
- Both voltmeters display the same voltage.

- 4 A voltmeter is used to measure the voltage (relative to ground) at various points around a circuit. The voltages at points X and Y are as shown in the diagrams. For circuits (a) and (b) determine the magnitude and direction (P or Q) of the current flowing through the 5 kΩ resistor and the 2 kΩ resistor.

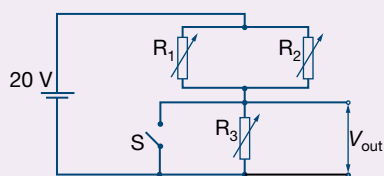


- 5 The following circuit is a simple voltage divider consisting of two variable resistors of resistance R_1 and R_2 . Copy and complete the table.



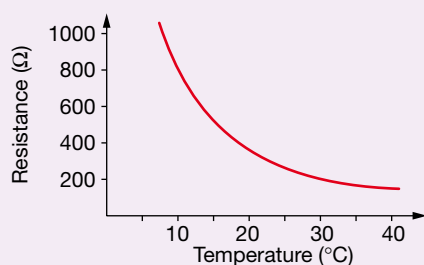
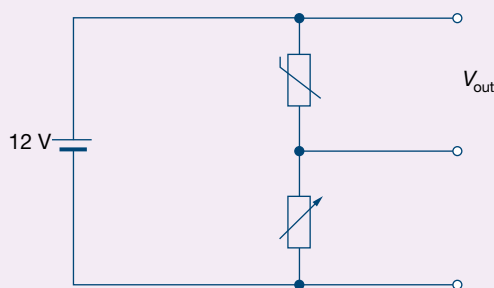
R_1 (Ω)	R_2 (Ω)	V_{out} (V)
1000		10
	1000	5.0
400	100	
900		2.0
2.0	3.0	

- 6 Three variable resistors, R_1 , R_2 and R_3 , are connected to a switch S. Determine the values of V_{out} in the table.



R_1 (Ω)	R_2 (Ω)	R_3 (Ω)	S	V_{out} (V)
200	200	100	open	
300	100	25	open	
50	50	75	open	
100	100	1000	open	
200	100	100	closed	

- 7 A portable refrigerator is controlled by the thermistor circuit shown. The control circuit is supplied by the 12 V battery of a car and it is required to keep the refrigerator temperature below 10°C . An output voltage of 4.0 V is required to turn the cooling system on. Using the characteristic curve of the thermistor shown, determine the required maximum value of the variable resistor.



The following information applies to questions 8–10. A manufacturer makes two different globes for a 12.0 V torch: one is rated at 0.50 W and the other at 1.00 W.

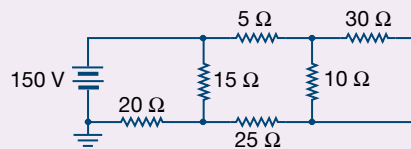
- 8 What is the resistance of each globe when it is operating correctly?
- 9 If the two globes are connected in parallel across a 12.0 V supply:
- which has the greater potential difference across it?

- which has the greater current through it?
- which glows more brightly?

- 10 If the two globes are connected in series across the 12.0 V supply:

- which has the greater potential difference across it?
- which has the greater current through it?
- which glows more brightly?

11



- a Which of the following combinations of resistors in the circuit shown are wired in series with each other?

- 30 Ω , 5 Ω
- 20 Ω , 5 Ω
- 20 Ω , 25 Ω
- 5 Ω , 10 Ω , 25 Ω
- 10 Ω , 15 Ω
- 5 Ω , 10 Ω , 25 Ω , 20 Ω

- b Which of the following combinations of resistors in the circuit are wired in parallel with each other?

- 30 Ω with 5 Ω
- 20 Ω with 15 Ω
- 30 Ω with 10 Ω
- 15 Ω with 10 Ω
- (5 Ω , 10 Ω) with (15 Ω , 25 Ω)
- (5 Ω , 10 Ω , 25 Ω) with 15 Ω

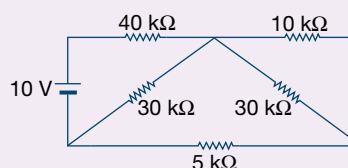
- c Which of the four resistors in the circuit shown has the least power dissipated through it?

- 20 Ω
- 15 Ω
- 25 Ω
- 5 Ω

- d Which of the four resistors in the circuit shown has the most power dissipated through it?

- 20 Ω
- 30 Ω
- 25 Ω
- 5 Ω

- 12 Consider the circuit shown.



- Determine the total resistance connected across the battery terminals.
- Determine the voltage across the 40 k Ω resistor.
- Determine the current through the 5 k Ω resistor.
- Determine the voltage across the 10 k Ω resistor.

4.2 Diodes

Ohmic and non-ohmic devices

The conductors and resistors that you have studied so far have been ohmic resistors; that is, they behave according to Ohm's law (see section 4.1). Ceramic resistors and most metal conductors at a constant temperature are ohmic resistors.

Electronic circuits use ceramic resistors to carry out various functions, such as controlling the voltage across a particular part of a circuit. Ceramic resistors, which are available in a range of resistance values, are sometimes called 'carbon-composite' resistors because they are made of powdered carbon mixed with a glue-like binder. The physical size of these resistors is not related to their resistance value. Rather, the size of a resistor relates to the electrical power that it can handle. Lower-resistance resistors are often physically larger than high-resistance ones, as they are designed to handle greater current and power loads. Resistors convert electrical energy to heat, and if a resistor gets too hot, it may not operate at the correct resistance value.

While ohmic devices obey Ohm's law, many other devices *change resistance* as the applied potential difference changes. Therefore, they are called non-ohmic devices. These include light globes, heating elements and diodes. A feature of these devices is that an increase in the potential difference across the component *does not produce a proportional increase* in the current flowing through the device.

Diodes

A diode is a non-ohmic electronic component that acts as a voltage-controlled switch. It allows electric current to flow freely in one direction, but (ideally) not in the opposite direction. Figure 4.12 represents a typical diode and its corresponding symbol. The band on the diode (and the perpendicular bar on the symbol) indicates which end is the cathode. The other end is called the anode. *For the diode to conduct, the cathode must be at a lower electrical potential than the anode.* The diode is said to be *forward biased*. The arrowhead in the symbol points in the direction in which conventional current can flow, with little resistance.

An ideal diode would have zero resistance when forward biased and an infinitely high resistance when reverse biased. In practice, a diode cannot display these ideal characteristics. Figure 4.13 shows the I versus V relationship for a typical diode. It is non-linear; therefore, we say a diode is a *non-ohmic device*. The right-hand side of the graph refers to a forward-biased diode. Note that when very small forward voltages are applied, the diode does not conduct. However, at a particular voltage called the *switch-on voltage* (V_s), or threshold voltage, the current flowing through the diode increases very rapidly. This corresponds to the resistance of the diode approaching zero and an effective short circuit existing between the leads of the diode—that is, the diode conducts. The value of V_s depends on the material that makes up the diode. For example, a silicon diode has a V_s of about 0.7 V and a germanium diode has a V_s of about 0.3 V.

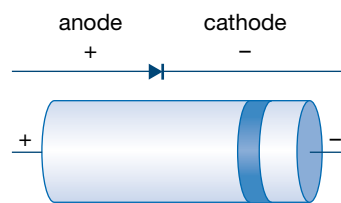


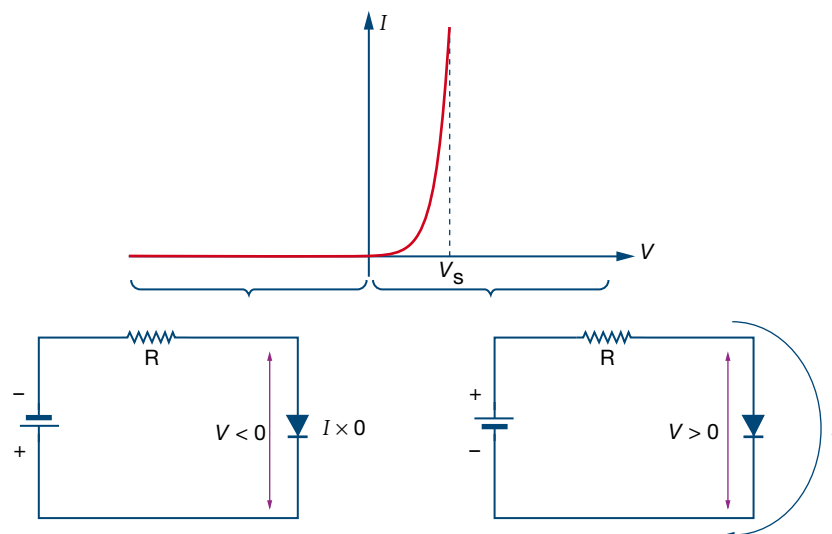
Figure 4.12 A typical diode and its circuit symbol.



PRACTICAL ACTIVITY 19

Characteristics of diodes

Figure 4.13 I – V characteristics of a diode. Current I passing through a diode as a function of the voltage.



When a diode is reverse biased, it will ideally prevent the flow of *any* current, no matter how large a potential difference is applied. In practice, there is a small leakage current which is, however, usually negligible in size—of the order of a few microamperes—and this has little effect on the circuit. A larger voltage, typically in excess of 50 V, will cause the diode to break down and current to flow.

When operating in forward-bias mode, the potential difference across an ideal diode will never rise noticeably above the switch-on (or threshold) voltage, V_s , regardless of how much current is flowing in the circuit. That is, a silicon diode should maintain a 0.7 V drop in potential regardless of the potential difference supplied. Increasing the voltage supplied to the circuit increases the circuit current and the potential drop across any resistor in series with the diode. In practice, for every diode there is a maximum forward current rating that should not be exceeded or overheating will occur, which damages the diode. The ability to maintain a constant potential difference, regardless of current, makes the diode a very useful component in a voltage divider circuit.

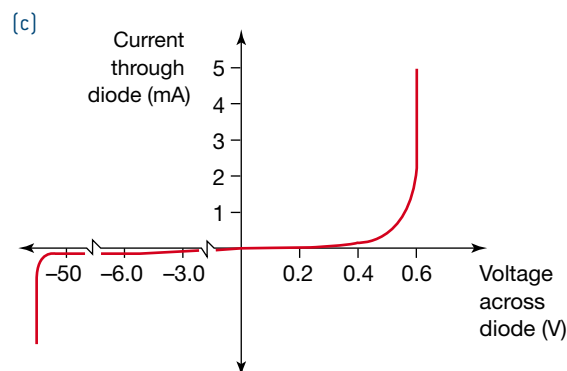
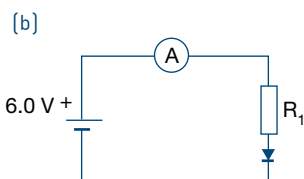
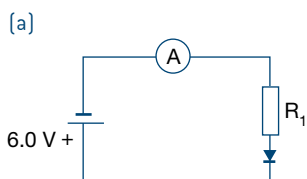
Physics file

The switch-on voltage of a diode is also called the 'threshold voltage'.

Worked example 4.2A

A student is investigating the current–voltage characteristics of a diode using the circuit shown in diagram (a). The I – V graph for this diode is illustrated in (c).

- a** Using the characteristics of the graph, describe what current the student could expect to measure using this circuit.



b From what material is the diode probably constructed?

The student then reverses the polarity of the battery, as shown in diagram (b), and measures the current as 4.5 mA.

c What is the potential difference across the diode?

d What is the resistance of resistor R_1 ? (Assume that the resistance of the milliammeter is negligible.)

Solution

a In the circuit shown, the diode is reverse biased. As a result very little current will flow. An ideal diode would let no current flow.

b The threshold voltage is around 0.6 V, which is characteristic of silicon diodes.

c The potential difference across the diode can be read directly from the I - V graph. At 4.5 mA the voltage across the diode will be 0.6 V. [The voltage will stay at this value as the current or battery voltage increases until reaching its maximum designed voltage.]

d We know that the battery voltage is 6.0 V, the voltage drop across the diode is 0.6 V and the current is 4.5 mA.

$$\text{Voltage across } R_1 = 6.0 - 0.6 = 5.4 \text{ V}$$

$$\text{Resistance } R_1 = \frac{V}{I} = \frac{5.4}{4.5 \times 10^{-3}} = 1200 \, \Omega$$

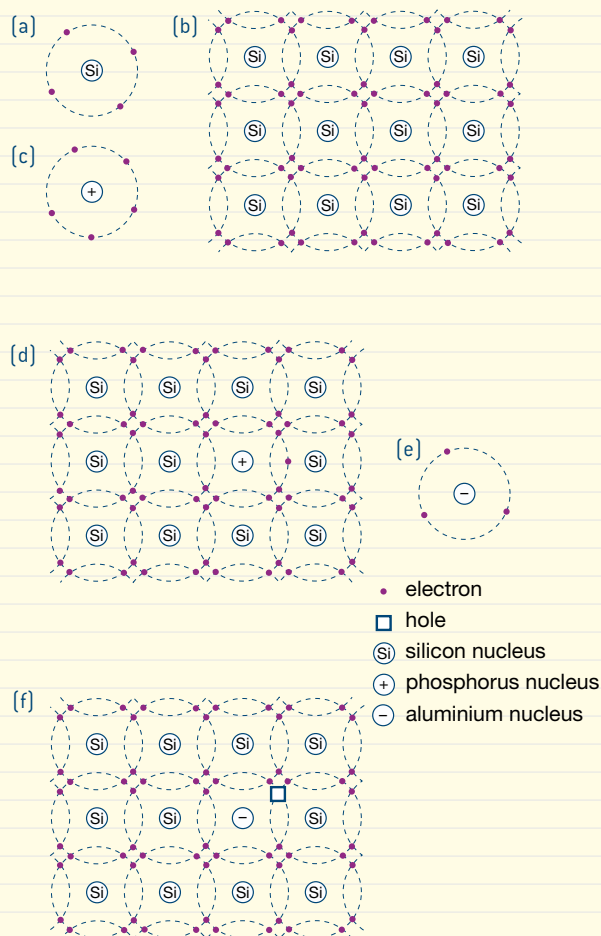
Physics in action

Inside a diode

Diodes are made up of highly ordered arrays (lattices) of atoms. The atoms (e.g. silicon or germanium) used in these lattices have four outer-shell electrons that are available for bonding (see Figure 4.14a). Each atom in the lattice shares its four electrons with its four nearest neighbours (covalent bonding) so that all atoms in the lattice appear to have a full shell of eight *shared* electrons, which is a stable configuration (see Figure 4.14b). Usually these electrons are held more tightly than in a conductor but not as tightly as in an insulator; hence, these lattices are called *semiconductors*. At room temperature, thermal atomic vibrations break a few of these strong bonds, generating a small number of mobile charges. In most diode circuits, the exceedingly small current (typically 10^{-9} A) resulting from these mobile charges can usually be ignored.

The conductivity of a semiconductor can be increased by a process called *doping*, which involves replacing some of the lattice atoms with atoms of other elements. For example, if we replace a few of the atoms in a silicon lattice with atoms that have one additional proton in the nucleus (e.g. phosphorus), then those dopant atoms will have one more electron per atom than is required for covalent bonding (see Figure 4.14c).

Figure 4.14 (a) A silicon atom. The silicon atom needs another four electrons to form a stable full shell. (b) A silicon crystal in which each Si atom 'sees' eight shared outer-shell electrons. (c) A phosphorus atom, which has one more electron in its outer shell (total of five) than silicon does. (d) A phosphorus-doped silicon crystal (an example of an n-type semiconductor). There is one extra loosely bound electron. (e) An aluminium atom, which has one less electron in its outer shell (total of three) than silicon does. (f) An aluminium-doped silicon crystal (an example of a p-type semiconductor). There is one loosely bound 'hole' (absence of an electron).



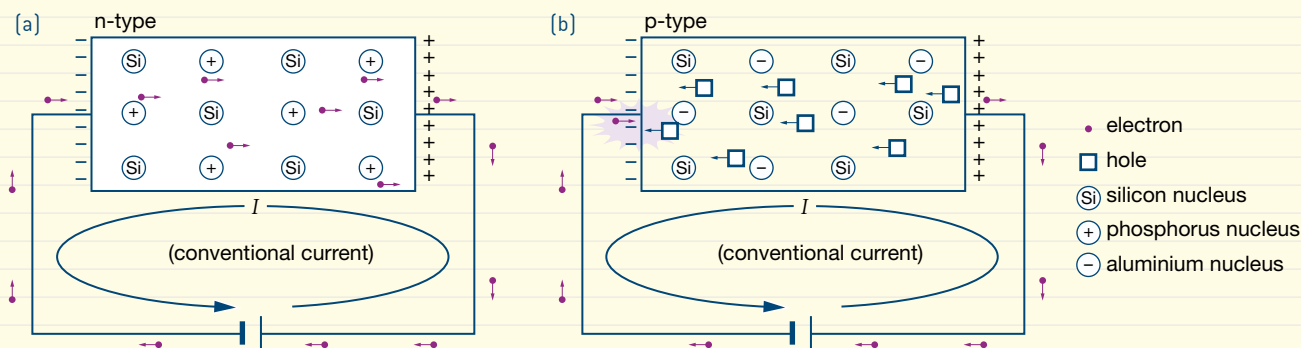


Figure 4.15 Doped semiconductors with voltage applied: (a) n-type semiconductor material; (b) p-type semiconductor material.

These electrons are very loosely bound, and only a very small amount of energy is needed to free them, creating mobile charges. This type of doping produces *n-type* (or negative-type) semiconductors, in which the spare electrons from the doping atoms can move relatively freely through the semiconductor (see Figure 4.14d).

If we replace a few of the atoms in a silicon lattice with atoms that have one less proton in the nucleus (e.g. aluminium) (see Figure 4.14e) this produces a *hole* (absence of an electron) in the lattice (see Figure 4.14f) into which an electron from an adjacent atom can easily move. The newly created hole in the adjacent atom can then be filled by an electron from another atom and so on. This type of doping results in *p-type* (or positive-type) semiconductors.

When a battery is connected across an n-type material, free electrons (in the semiconductor) flow towards the positive end of the semiconductor (see Figure 4.15a). In p-type material, connecting a battery causes (in effect) a flow of holes towards the negative end of the semiconductor (and therefore a conventional current in the same direction). At the negative end of the semiconductor, free electrons from the battery 'fill' the holes. Meanwhile at the positive end, electrons flow into the wire, creating holes in the p-type semiconductor material. These holes will drift through the p-type material, establishing a continuous current for as long as the battery remains connected (see Figure 4.15b).

A diode is made by creating a junction between a piece of n-type and a piece of p-type semiconductor material. Here we shall look at a silicon diode. Some of these electrons will diffuse across the junction from the n-type into the p-type material and fill some of the vacant holes. This process is called recombination. The same thing happens when holes diffuse from the p-type material into the n-type material. This produces a region around the junction that has very few mobile electrons and holes. This region is called the *depletion region*, since it is depleted of free charge.

The depletion of mobile charges around the junction means that in the n-type part of the depletion layer, each phosphorus atom now has a net positive charge (i.e. 15 protons and 14 electrons), whereas in the p-type part, each aluminium atom has a net negative charge (i.e. 13 protons and 14 electrons). This produces a net positive charge in the n-type material near the junction and a net negative charge in the p-type material. This charge separation results in an electric field and an electrical potential barrier (V_0) across the junction that opposes the diffusion of any more mobile charge (see Figure 4.16a). For a silicon pn junction V_0 is approximately 0.65 V. Other semiconductors will have different barrier potentials.

Applying an external battery so that the positive terminal of the battery is connected to the n side of the diode and the negative terminal is connected to the p side is called *reverse biasing* the diode. The external reverse-biased potential difference across the diode increases the pn junction's effective potential barrier. It also widens the depletion layer by attracting both holes and electrons away from the pn junction, broadening this non-conducting region. Hence, the reverse bias ensures that no current will flow through the diode (see Figure 4.16b).

When an external battery is applied so that the positive terminal of the battery is connected to the p side of the diode and the negative terminal is connected to the n side, this is called *forward biasing* the diode. Applying an external forward bias decreases the effective potential barrier and narrows the pn depletion layer. This means that electrons migrate from the n-type region through the depletion region into the p-type region where they can fill holes (recombination). Holes migrate in the opposite direction and recombine with electrons. Thus, holes are attracted towards the negative terminal and electrons are attracted to the positive terminal (Figure 4.16c). Holes are continuously produced at the positive end of the p-type material and electrons enter the material at the negative end of the n-type material. The continuous flow of charge is seen as a flow of current. The size of the current flowing through the diode strongly depends on the forward bias voltage applied across the diode.

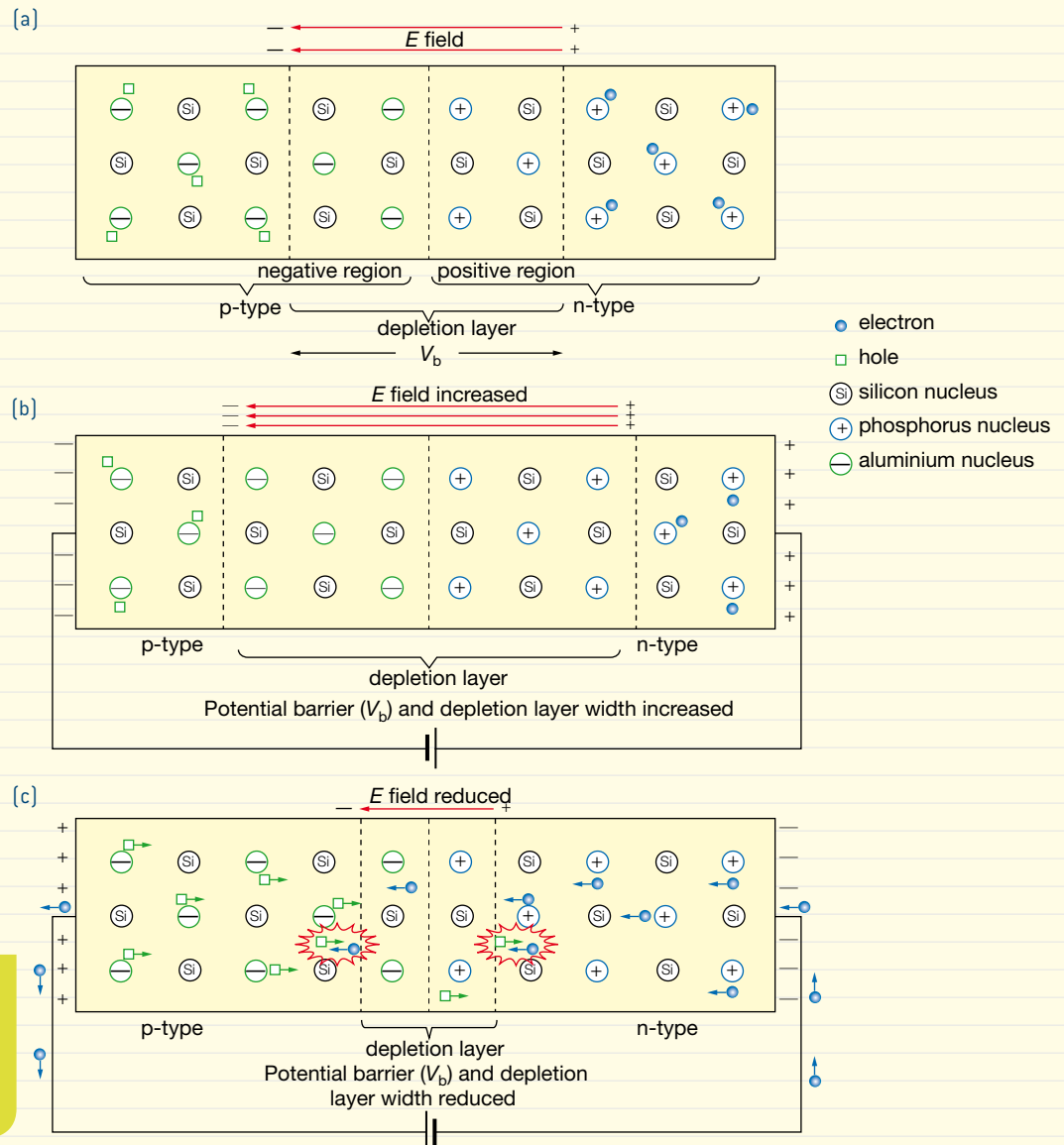
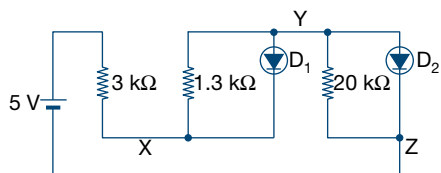


Figure 4.16 (a) An unbiased pn semiconductor junction. (b) A reverse-biased pn semiconductor junction. (c) A forward-biased pn semiconductor junction.

Worked example 4.2A

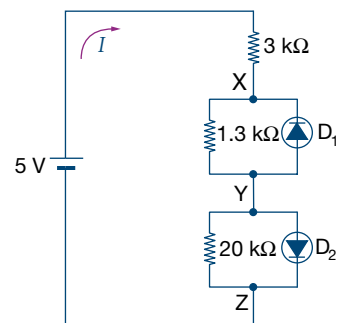
For the following circuit, determine the voltage across D_1 and the current flowing through D_2 . Assume that D_1 and D_2 are both silicon diodes.



Solution

Redraw the circuit [shown at right].

Note that D_1 is reverse biased so its resistance is large [i.e. $R_{D_1} \gg 1.3 \text{ k}\Omega$]. Thus, the effective resistance between X and Y is 1.3 kΩ. Note that D_2 is forward biased so its resistance is



small (i.e. $R_{D_2} \ll 20 \text{ k}\Omega$). Thus we can ignore the effect of the $20 \text{ k}\Omega$ resistor, which is in parallel with D_2 .

D_2 switch-on voltage is 0.7 V .

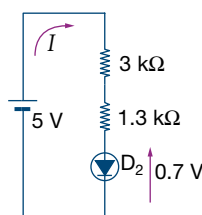
Therefore, voltage drop across rest of circuit is 4.3 V . Current flowing from battery is:

$$\begin{aligned} I &= \frac{V}{R_t} \\ &= \frac{4.3}{4.3 \times 10^3} \\ &= 1 \times 10^{-3} \text{ A} \\ &= 1 \text{ mA} \end{aligned}$$

Voltage across D_1 = voltage across $1.3 \text{ k}\Omega$ resistor

$$\begin{aligned} V_{D_2} &= 1 \times 10^{-3} \times 1.3 \times 10^3 \\ &= 1.3 \text{ V} \end{aligned}$$

Current through $D_2 = 1 \text{ mA}$



Physics in action

The cathode ray oscilloscope

For time-varying voltage signals it can be useful to be able to 'graph' the exact shape of the signal or its timing relative to other signals. In these cases, a cathode ray oscilloscope (CRO) is used. A CRO (which is often also simply called an oscilloscope) effectively takes a 'snapshot' of any time-varying voltage signal.

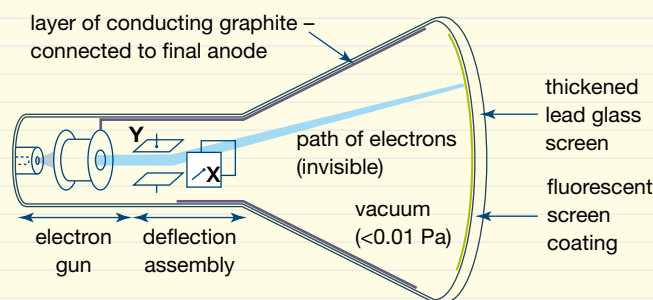


Figure 4.17 The cathode ray oscilloscope.

The oscilloscope converts the electrical signal (voltage) into a deflection of a beam of electrons. The electron beam excites a phosphor screen to produce a visible light pattern that can be sensed by our eyes (see Figure 4.17).

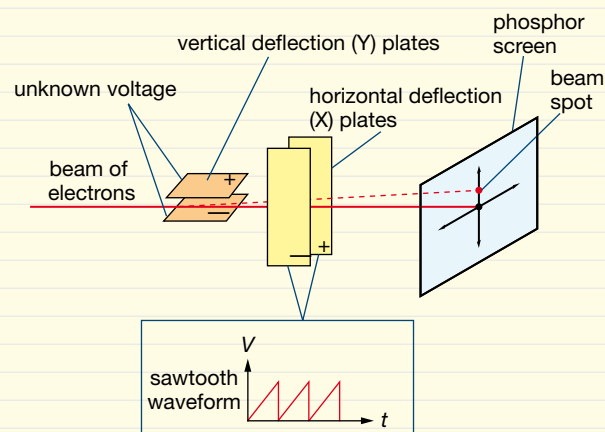


Figure 4.18 Deflection of an electron beam in the oscilloscope.



Figure 4.19 A typical two-channel oscilloscope.

The beam passes with a constant speed through two sets of deflection plates (as shown in Figure 4.18). These plates allow the beam to be deflected in both the vertical and horizontal directions before hitting the phosphor screen.

Hence, the voltage applied to the Y (or X) set of plates is proportional to the vertical (or horizontal) displacement of the beam spot on the phosphor screen. The electrical signal to be measured (signal voltage) is applied to the Y plates. To observe the time-varying nature in the signal being measured, the beam spot must also be swept repeatedly in a horizontal (X) direction with a constant speed.

Once synchronised, accurate measurements (both time and voltage) can then be made from the stationary pattern displayed. The screen of the oscilloscope can be likened to a sheet of graph paper. The voltage being measured is displayed along the vertical axis, and time is along the horizontal axis.

The time (time/div) and voltage (volts/div) scale knobs on the CRO can be used to change the *scale* of the displayed signal in both the horizontal and vertical direction. Note that the oscilloscope cannot change the actual signal itself, only how that signal is displayed. We often use an oscilloscope to display periodic voltage signals that oscillate too rapidly to be seen by the human eye.



4.2 summary

Diodes

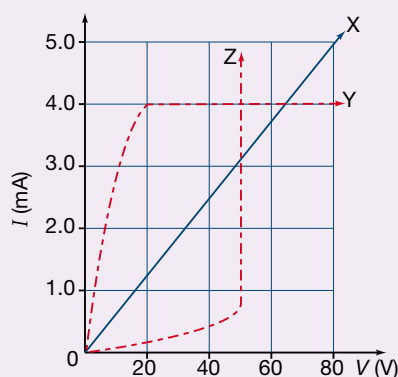
- A device in which the ratio of *current to potential difference* changes is called non-ohmic.
- Diodes are non-ohmic devices used to control the direction of current flow within a circuit.
- A diode conducts current only in one direction, i.e. when it is forward biased. When a diode is forward biased, a minimum threshold potential must be applied before the diode will begin to conduct. The threshold potential depends on the material from which the diode is made.
- When forward biased and conducting, an ideal diode maintains a constant potential difference.
- Commonly, silicon diodes are used that are characterised by a threshold voltage, V_s , of 0.7 V.
- A characteristic curve will show the operating specifications of a diode. It is commonly assumed that when an ideal diode is reverse biased, negligible current will flow through it and it can be considered to have infinite resistance (effectively an open circuit).



4.2 questions

Diodes

- 1 An ideal diode with a switch-on voltage of 0.65 V is placed in series with a 4000 Ω resistor. A DC supply voltage of 6.0 V is applied. Determine the size of the current through the resistor if:
 - a the diode is reverse biased
 - b the diode is forward biased.
- 2 The current–voltage characteristics for three different circuit devices X, Y and Z are shown in the graph.



- a Which of these devices:
 - i is non-ohmic?
 - ii has a constant resistance?
 - b Determine the resistance(s) of the ohmic device(s).
 - c All of the three devices are now connected in parallel across a battery. The current in device Z is 2.0 mA. Determine the current flowing in devices X and Y.
- 3 a When a silicon diode is forward biased with a voltage of 0.7 V, a current I_1 flows through it. If the forward-biased voltage is halved, what is the new current, I_2 , flowing through the diode? Choose from A–E below.

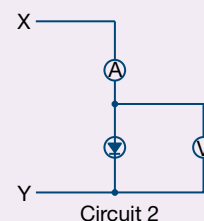
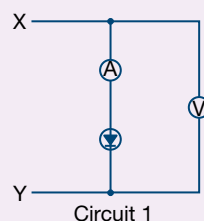
- b When a silicon diode is reverse biased with a voltage of 5 V, a very small thermal leakage current, I_1 , flows through it. If the reverse-biased voltage is doubled, what is the new current, I_2 , flowing through the diode? Choose from A–E.

- A $I_2 < I_1/2$
- B $I_2 \approx I_1$
- C $I_2 = 2I_1$
- D $I_2 = I_1/2$
- E $I_2 > 2I_1$

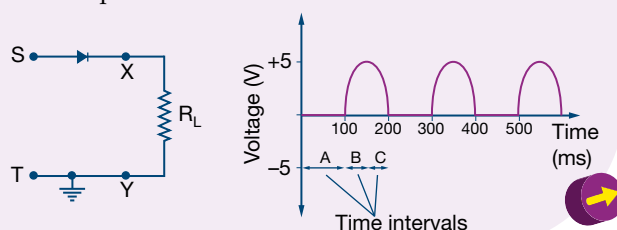
- 4 Which of the following circuits would you use to measure the V – I characteristics of a:

- a forward-biased diode, i.e. $V_x > V_y$?
- b reverse-biased diode, i.e. $V_x < V_y$?

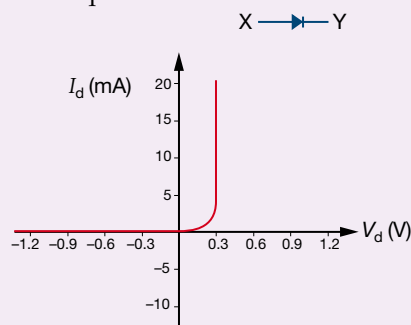
Give reasons for your answers. Remember that to measure the I – V characteristics we need to measure the voltage across, and the current through, the diode as accurately as possible.



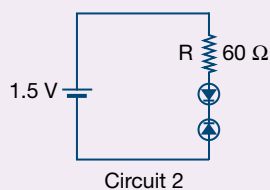
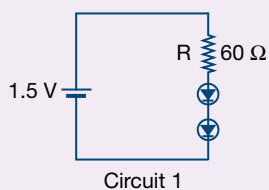
- 5 The graph shows the voltage (v_{xy}) across the load resistor plotted as a function of time.



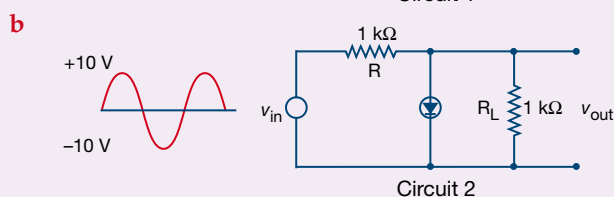
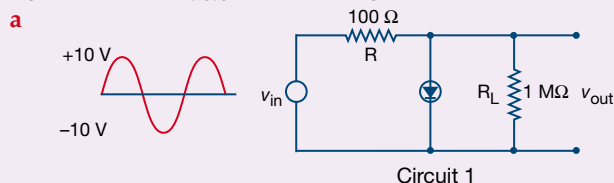
- a i** During which time interval(s) (A, B or C) is the diode forward biased?
- ii** During which time interval(s) (A, B or C) is the diode reverse biased?
- b** Draw the display that would be observed for the voltage between S and T (V_{ST}), assuming it is symmetric about zero and a silicon diode is used.
- 6** The following I - V graph applies to all diodes depicted in this question.



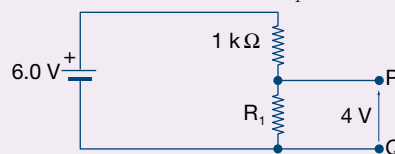
- a i** Explain the meaning of 'switch-on voltage'.
- ii** What is the switch-on voltage for this diode?
- iii** From what material is this diode most likely made?
- iv** If the diode is reverse biased, is $V_X > V_Y$ or vice versa?
- v** If the diode is forward biased, does conventional current flow from X to Y or vice versa?
- b** Determine the power dissipated in each of the two $60\ \Omega$ resistors (shown in circuits 1 and 2).



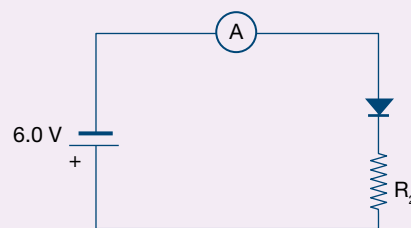
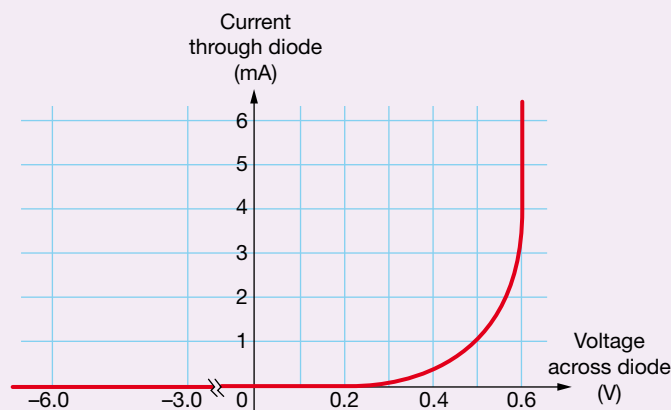
- 7** Sketch the output waveforms (v_{out}) for each of the following circuits. Assume the silicon diode has a switch-on voltage $V_S = 0.7\text{ V}$. Assume $R_{diode} \rightarrow 0\ \Omega$ for $V_S \geq 0.7\text{ V}$ and $R_{diode} \rightarrow \infty\ \Omega$ for $V_S < 0.7\text{ V}$.



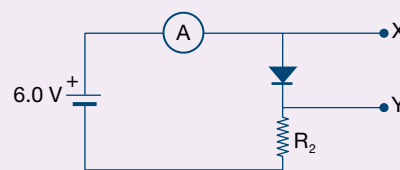
- 8** Two resistors are used to make a voltage divider to use with a 6.0 V battery, as shown below, in order to obtain a 4 V output between P and Q. What is the resistance of the resistor R_1 ?



The following information applies to questions 9 and 10. A diode with the characteristics shown on the graph is used in the circuit also shown below.



- 9** Using the characteristics of the diode, explain why the current in the circuit is extremely small.
- 10** The circuit is now adjusted as shown.



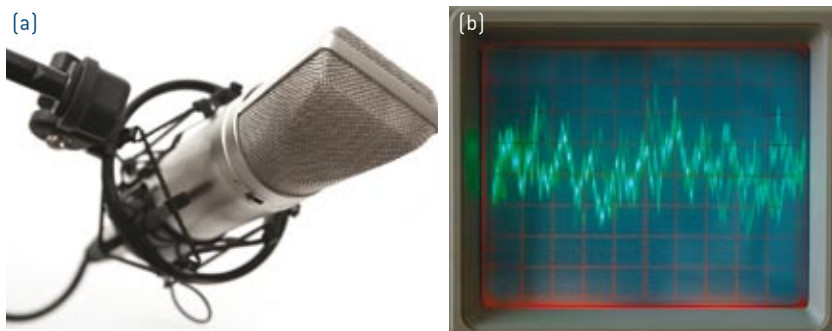
The current flowing is now 4.5 mA .

- a** What adjustment has been made to the circuit when compared with question 9?
- b** What is the voltage, V_{XY} , across the diode?
- c** What is the resistance of R_2 ? The resistance of the ammeter in the circuit can be ignored.

4.3 Amplification

Voltage amplifiers

Some voltages in electronic circuits are time-varying in nature and have a small amplitude. These small, varying voltages are not intended to provide energy to a circuit, but to carry information such as radio transmission signals. They are referred to specifically as *signal voltages*. They have become an important part of electronics as we get better at miniaturising components and detecting small currents.



A microphone is an example of an electronic device that produces small signal voltages. A microphone (Figure 4.20a) converts sound waves into an equivalent electrical signal: a small, varying voltage whose amplitude may be only a fraction of a volt. The variation of the signal carries information about the frequency and tonal characteristics of the sound. Although an oscilloscope or datalogger can detect this sort of signal voltage, the signal does not have sufficient energy to activate many other electrical devices—it often needs to be amplified to make the information within it more accessible. In the case of a microphone, an *amplifier* magnifies the voltages to levels (typically tens of volts) that can be fed into an $8\ \Omega$ loudspeaker to produce an audible sound. Ideally, an amplifier produces an output signal with an identical wave shape to the input signal but with larger amplitude, as shown in Figure 4.20b.

The operation of a voltage amplifier depends largely on transistors, which are not covered in this course. An amplifier essentially uses a small voltage variation at its input (V_{in}) to control a large voltage variation at its output (V_{out}). The amplifier requires its own power source before it can amplify the signal.

This is an important point. An amplifier circuit will have two inputs. One input is the small signal voltage (usually AC) that is going to be amplified. The other input is the relatively larger DC supply, which, of course, is the source of the extra energy that is provided during amplification (see Figure 4.20c).

Normally, we require the output voltage to be directly proportional to the input. Consequently, a graph of V_{out} as a function of V_{in} would be a straight line and the amplifier is then said to be *linear*. A practical problem with voltage amplifiers is that, at least to a small extent, they fail to meet these requirements precisely. The shape of the amplified signal can change if the input signal has a large amplitude, resulting in *non-linear amplification*. In the case of a sound amplifier, this can change the quality of the sound, and in an extreme case the output may become badly distorted.

Physics file

We tend to analyse functioning amplifiers by looking at the peak-to-peak values of the input and output AC signals.

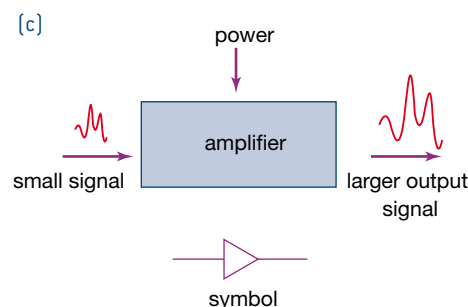


Figure 4.20 (a) A microphone converts sound waves to a small, rapidly varying signal voltage. (b) An oscilloscope can be used to view this signal. (c) To recreate the sound, a voltage amplifier is required to increase it to a usable amplitude. Amplifiers require a power source to provide the electrical energy to magnify the signal voltage.

Physics file

Transistors were developed during the late 1940s to replace the bulky glass valves then being used to amplify signal voltages. By the mid 1950s, transistor sales were already in the billion-dollar range, and today, the information revolution powered by cable modems, high-speed digital networks and worldwide communication networks depends upon this relatively simple device. The very high-density integrated circuit 'chips' used in modern computers can contain the equivalent of millions of tiny transistors.

Voltage gain

Simple amplifying circuits use just one transistor. The input voltage is effectively multiplied by a selected factor. The simplest single-stage amplifiers might multiply an input signal by a factor of ten, or a factor of 400, for example.

In most real amplifying circuits, multiple transistors are used, each one repeatedly multiplying the output of the previous stage of amplification. In this course, you need only show an understanding of the function of amplifiers overall, and not the details of its components.

When the output voltage for a typical inverting amplifier is plotted as a function of the input voltage, a graph, like that shown in Figure 4.21, is obtained. The following points can be seen from studying the graph.

- The value of V_{out} is determined by V_{in} . For this *inverting* transistor amplifier, an increase in V_{in} causes a decrease in V_{out} . The maximum value for V_{out} is the supply voltage for the amplifier. The graph indicates that a range of values for V_{in} will give this value for the output voltage. This is a non-linear region of the graph. For these values, the transistor is said to be in *cut-off* mode. Similarly, a high value for V_{in} causes a low value of V_{out} . This minimum value is close to zero. The graph indicates that a range of values for V_{in} will give this value for the output voltage. Again, this is a non-linear region of the graph. For these values, the transistor is said to be in *saturation* mode.
- Distortion-free amplification will occur in the region where the graph is linear. To correctly amplify a small time-varying signal, the amplifier needs to be operated so that the input voltage varies about the middle of the linear region, well away from the 'cut-off' and 'saturation' non-linear regions. This means that a constant offset voltage needs to be added to the varying input signal so that V_{in} is in the middle of the linear amplification region. This offset voltage is known as the *bias voltage* of the amplifier.

The amplification *factor* of the amplifier is referred to as the *gain*. The *voltage gain* (A_v) is the ratio of V_{out} to V_{in} (i.e. $A_v = V_{out}/V_{in}$), and it tells us by how much the input signal is scaled up at the output. Voltage gain is a ratio with no units.

In practice, the amplifier circuit is constructed so that the input time-varying signal voltage (v_{in}) is *added* to the bias voltage. When there is no AC input, the circuit therefore has an input of only the bias voltage. This allows the greatest peak-to-peak input signal voltage to be amplified without distortion (see Figure 4.22).

The voltage gain for a *time-varying input signal*, A_v , is the ratio of Δv_{out} to Δv_{in} .

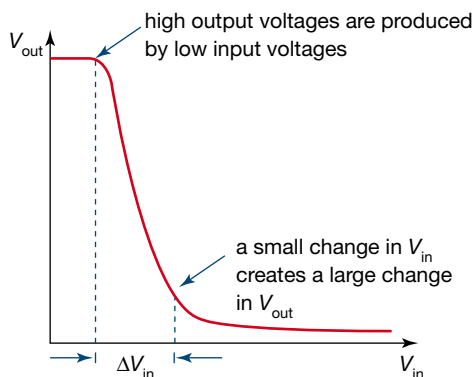


Figure 4.21 The graph of V_{out} versus V_{in} for a voltage amplifier. A small variation in voltage at the input will produce a large variation in voltage at the output. The output changes from a high value to a low value for a very small change in the input voltage.

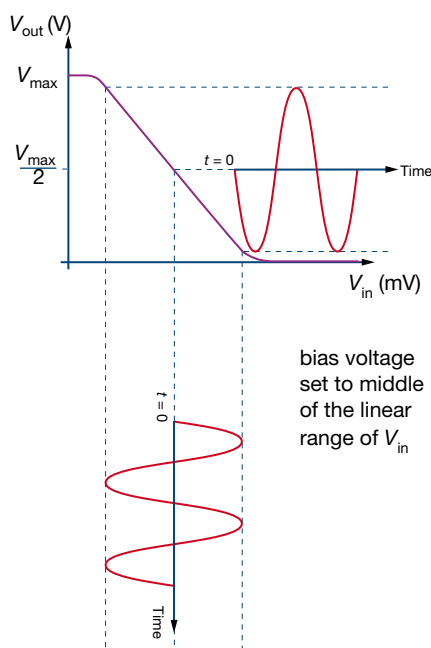


Figure 4.22 Graph of V_{out} versus V_{in} for a correctly biased voltage amplifier. A large time-varying input signal can be amplified without distortion.



VOLTAGE GAIN is given by:

$$A_v = \frac{\Delta v_{out}}{\Delta v_{in}}$$

where A_v = voltage gain (no unit)

Δv_{out} = range over which the output voltage varies (V)

Δv_{in} = range over which the (AC) input voltage varies (V)

Gain, A_v , is also given by the gradient of the Δv_{out} vs. Δv_{in} graph. A negative value of A_v is often stated for inverting amplifiers.

Amplifier clipping

Figure 4.23a represents the voltage gain of a complex *inverting* voltage amplifier that can amplify both positive and negative input signals. The maximum output voltage of the amplifier has a peak of +6 V. This is first achieved when the input voltage falls to -0.1 V. As the input voltage decreases below -0.1 V, the output of the amplifier will remain at +6 V. The amplifier cannot produce a higher voltage. It has reached its *cut-off voltage*. The output has been clipped at its maximum output of 6 V. Similarly, the minimum output voltage of the amplifier is -6 V, and this is first achieved when the input voltage rises to $+0.1$ V. As the input voltage rises above $+0.1$ V, the output remains at -6 V and we say that the amplifier has reached *saturation*. This form of amplifier distortion is known as clipping. It occurs when the input signal is outside the amplifier's input limits. For input voltages between -0.1 and $+0.1$ V we have *linear* amplification resulting in output voltages between $+6$ and -6 V without distortion.

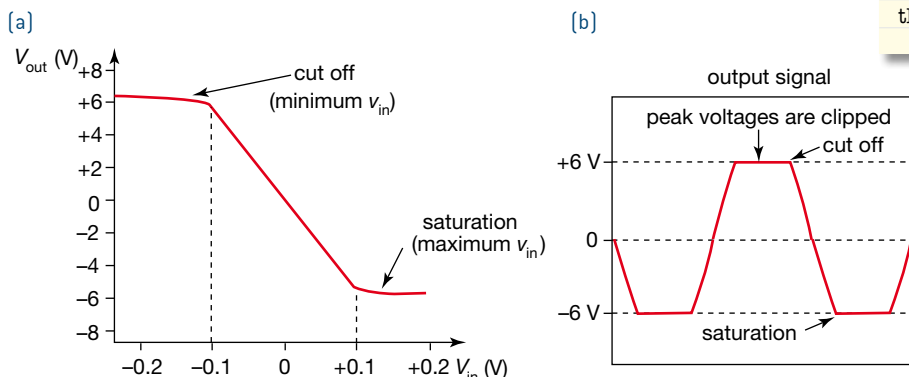
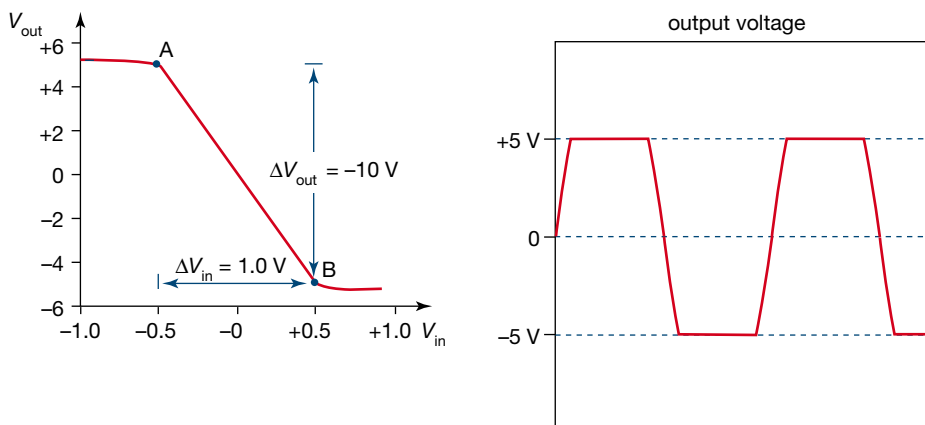


Figure 4.23 The operational characteristics of a voltage amplifier will limit the output voltage range. In this example (a) the peak output voltage of ± 6 V is achieved at an input voltage of ± 0.1 V, after which (b) the signal is *clipped* and distorted.

Figure 4.23b shows an example of amplification distortion. When the input voltage goes beyond the -0.1 and $+0.1$ V limit, the output voltage remains fixed at ± 6 V, and the output will be clipped as shown. If this were an audio amplifier, the sound output would be badly distorted.

Worked example 4.3A

An audio amplifier includes a voltage amplifier to amplify the AC signal input voltage from a small microphone. The voltage characteristics are shown in the diagram.



Physics file

A single-stage (one-transistor) amplifier will normally invert an AC signal (i.e. change its sign from positive to negative or vice versa), so if it is important that a signal not be inverted, a two-stage (two-transistor) amplifier must be used. Here, the input of a second transistor amplifier is connected to the output of the first amplifier. The overall gain for a two-stage amplifier will be the product (i.e. multiplication) of the gain for each individual transistor:

$$\Delta V_{\text{out}} = (A_{v_1} \times A_{v_2}) \Delta V_{\text{in}}$$

Since the individual gains (A_{v_1} and A_{v_2}) are both negative numbers, their product ($A_{v_1} \times A_{v_2}$) is positive and hence the amplifier is non-inverting.

Physics file

Throughout this chapter, DC voltages and currents are usually represented by UPPER-CASE characters (V and I) and AC voltages and currents are usually represented by lower-case characters (v and i).

Physics file

When an amplifier circuit is set up with the correct resistor values and DC voltages, ready to receive an AC signal, we call this the 'quiescent point'. Do not assume that amplifier circuits would have a DC current flowing at all times, even when there was no input to be amplified, as this would waste lots of energy! Most real amplifier circuits do not constantly draw current; instead they are designed to switch on when an input signal is detected.



PRACTICAL ACTIVITY 20

Transistors as amplifiers

Physics in action

Transistor amplifier circuits

An essential component of many amplifier circuits is the npn transistor. A diagram of the transistor and its electronic symbol is shown in Figure 4.24. An npn transistor is able to respond to the voltages that will be applied to it in a very particular way. As you can see from Figure 4.24, transistors have three connecting leads—collector, base and emitter leads.

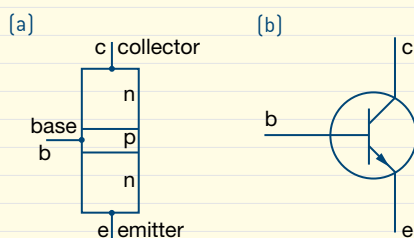


Figure 4.24 (a) npn junctions and (b) npn transistor circuit symbol.

A transistor is essentially a special type of switch. When it is placed appropriately in a circuit and a very small current is sent through the base-emitter path of the transistor, this can 'switch-on' a much larger current flowing from the collector to the emitter. This is because the flow of current across the base-emitter junction affects the conductance of the entire transistor, allowing a collector-emitter current pathway to be created (see Figure 4.25).

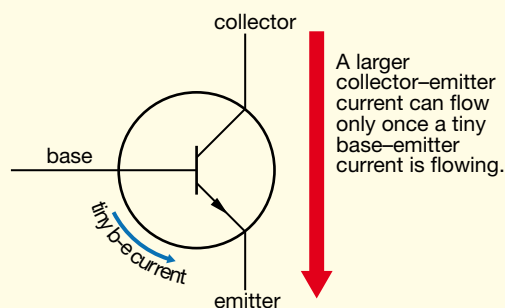


Figure 4.25 For the base-emitter pathway to conduct the base must be approximately 0.65 V higher potential than the emitter. Another [DC] power supply is the source of the larger collector-emitter current.

Think about the values of the potential required to create the base-emitter current. The base-emitter junction behaves very like a diode (studied in the previous section). We can say that the base-emitter junction needs to be forward biased for current to flow; that is, the base must be at a higher voltage (potential) than the emitter. A minimum switch-on voltage must be applied for the base-emitter junction to conduct. Just like a diode, once conducting the base-emitter junction will maintain an almost steady potential. A conducting base-emitter junction of a silicon transistor can be assumed to have a voltage across it that varies very slightly around a value of 0.65 V.

- What is the voltage gain (for a time-varying signal) of this amplifier?
- The amplifier can be described as an inverting amplifier. Why?
- The microphone produces a varying input voltage peaking at 0.5 V. What is the comparable output voltage?
- The input signal voltage is increased to a peak of 1.0 V. Assuming that the input signal is sinusoidal, sketch the likely output signal and describe the effect on the output.

Solution

a Voltage gain: $A_v = \frac{\Delta v_{\text{out}}}{\Delta v_{\text{in}}}$

Choosing points A and B from the graph to give as large a range as possible:

$$\Delta v_{\text{out}} = -10 \text{ V}, \Delta v_{\text{in}} = 1.0 \text{ V}$$

$$\text{and voltage gain } A_v = -\frac{10}{1.0} = -10$$

There are no units, since this is a ratio. Note that because the slope of the linear portion of the graph is negative, the gain of the amplifier is negative.

- b This voltage amplifier is described as an inverting amplifier because v_{out} is directly proportional to v_{in} but of opposite sign.

- c From the graph, an input voltage of 0.5 V corresponds to an output voltage of -5 V. We could also find this by using the gain calculated in part a:

$$\begin{aligned} V_{\text{out}} &= \text{gain} \times V_{\text{in}} \\ &= -10 \times 0.5 \\ &= -5.0 \text{ V} \end{aligned}$$

- d Above an input voltage of $\pm 0.5 \text{ V}$ the signal will be clipped. The output will peak at $\pm 5 \text{ V}$, as in the trace shown.

Since a very small current flowing in the base-emitter pathway of a transistor can result in a much larger current being conducted from the collector to the emitter lead of the transistor, there must be a source of this electrical energy. Later we will see how a DC power supply is used in the overall amplifier circuit. Note that when the transistor is operating, the current flowing out of the emitter lead of the transistor (I_e) is made up of both the base-emitter current (I_b) and the collector-emitter current (I_c). We say that $I_b + I_c = I_e$. Since the collector-emitter current (I_c), is often 100 or more times larger than the base-emitter current (I_b) we can say that $I_c \gg I_b$; therefore, $I_e \approx I_c$.

When an amplifier circuit is used correctly, a small change in the varying input *voltage* (applied across the base-emitter) results in small fluctuations in the base-emitter *current*. In turn, this results in large corresponding fluctuations in the collector-emitter current. The ratio of the magnitudes

of the collector current and the base current is thought of as the current gain (A_i) of the transistor. When a transistor is operating within its limits, there will be a consistent current gain and the collector current will always be proportional to the base current. This means that a varying input signal can be faithfully reproduced with no alteration to the frequency of variation.

However, there are limits to the operation of the transistor. Once the saturation voltage of the base-emitter junction is reached, no further increase in the base current or collector current can occur. Further increases in the base current will not be linearly amplified. We say that transistor **saturation** has occurred. Alternatively, if the base-emitter voltage drops below the minimum switch-on voltage required to activate the diode-like base-emitter junction, then the transistor stops conducting a base current and therefore no collector current will flow either. We say that transistor **cut-off** has occurred.

Physics in action

Transistor action in an npn BJT: How it works!

A common transistor in electronic amplifying circuits is the bipolar junction transistor (BJT), which is formed from three layers of doped semiconductor material. One configuration of BJT has a very thin layer of p-type material sandwiched between two layers of n-type material.

For normal operation with an npn transistor, the base is at a slightly higher potential than the emitter (~0.7 V for a silicon npn BJT). Hence, the pn 'diode' between base and emitter is forward biased (since the effective barrier potential is reduced). As we saw with the diode, a small change in V_{be} has a large effect on the number of electrons diffusing from the n-type emitter region into the p-type base region. In a normal pn diode, this would result in a large number of electrons flowing into the emitter and out of the base. In a transistor, because the base is very thin and lightly doped, very few of the electrons from the emitter 'fill' the small number of holes in the base and the base current is only a very small fraction of the emitter current. So where do all the electrons in the base go? The collector is at a much more positive potential than the base (since the pn 'diode' between the base and the collector is strongly reverse biased). This means that most of the mobile electrons that have diffused into the n-type base region are strongly attracted to the large positive potential of the p-type collector region. These electrons are quickly swept across the narrow base into the collector region by the very strong electric field. Reverse biasing the base-collector junction means that most of the electrons 'emitted' by the emitter into the base are 'collected' by the collector (see Figure 4.26).

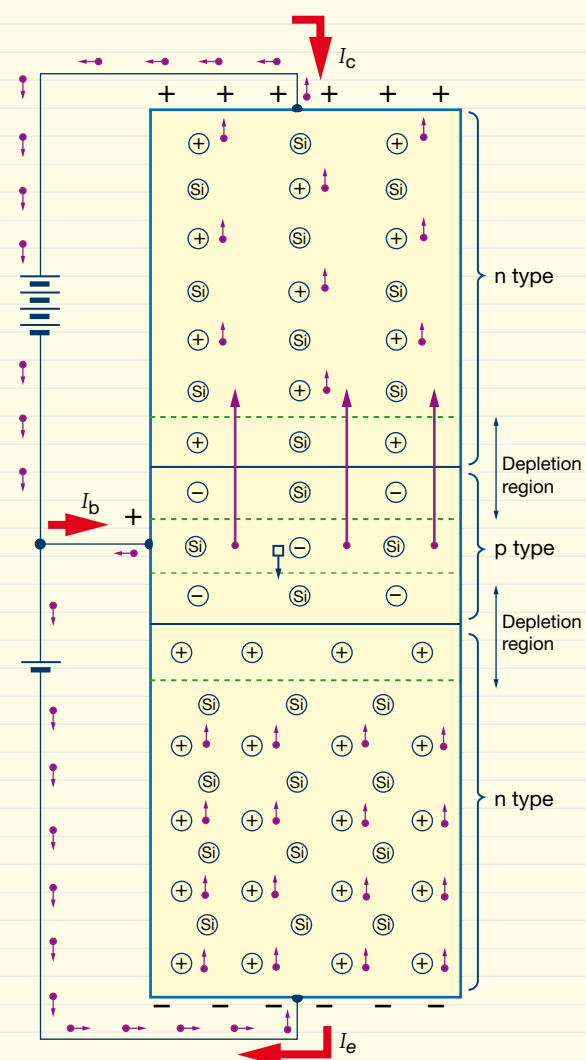


Figure 4.26 An npn transistor, showing npn semiconductor regions, depletion layers, biasing, flow of electrons and conventional current.



4.3 summary

Amplification

- Amplifier circuits convert a small varying AC voltage into a large varying voltage with the same characteristics, by utilising the amplifier's additional power supply.
- Amplifier circuits have a limit to the range of input voltages that can be amplified *linearly*. If this range is exceeded, clipping occurs and the output is distorted.
- When an amplifier is operating in the *linear* gain region, the voltage gain is consistent. A negative gain indicates an inverting amplifier.

- Voltage gain is given by:

$$A_v = \frac{\Delta v_{\text{out}}}{\Delta v_{\text{in}}} = \text{gradient of } v_{\text{out}} \text{ vs. } v_{\text{in}} \text{ graph}$$

- Biasing a transistor circuit means setting it up so that when there is no AC input, the output voltage is the value in the middle of its possible output range. This allows for the greatest possible variation in v_{in} , without distortion occurring.



4.3 questions

Amplification

- 1 Briefly explain the following terms.

- a linear amplification
- b voltage gain
- c clipping
- d current gain
- e inverting amplifier

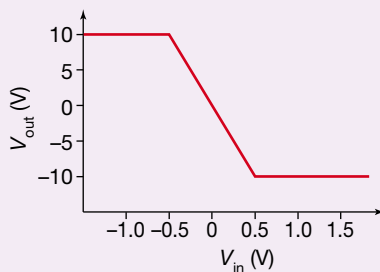
- 2 An amplifier circuit has been correctly biased and can produce a voltage gain of 200. A small *zero-centred* AC 40 mV peak-to-peak voltage signal is fed onto the amplifier.

- a Determine the peak-to-peak variation in the output voltage (Δv_{out}).
- b The input signal is now increased and clipping is known to occur. Explain the effect on the output voltage.

- 3 What is meant by the term 'amplifier biasing'?

The following information applies to questions 4–6.

The graph shows the characteristics of a voltage amplifier.

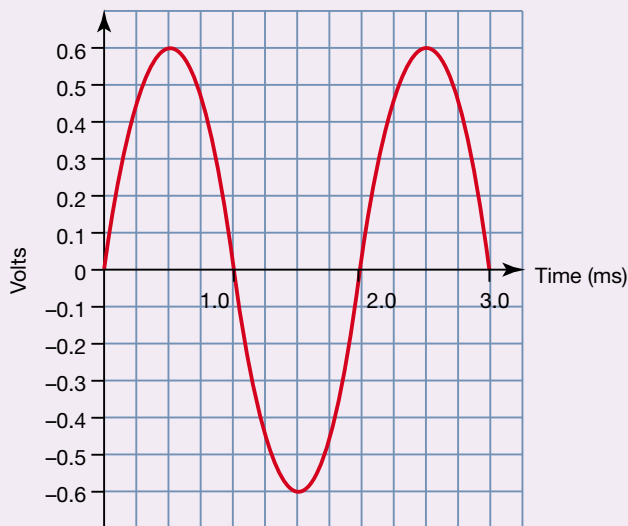


- 4 a Explain why the amplifier produces a constant output voltage over a range of input voltages.
b What is the maximum peak signal voltage that can be amplified without distortion?
c Calculate the voltage gain of the amplifier.
- 5 Complete the following table by calculating the range of output voltages for the amplifier for the corresponding sinusoidal input voltages with ranges as shown.

Range of input voltage	Range of output voltage
100 to 200 mV	
0.25 to 0.50 V	
1.0 to 1.50 V	
20 to -20 mV	
0.80 to -0.80V	

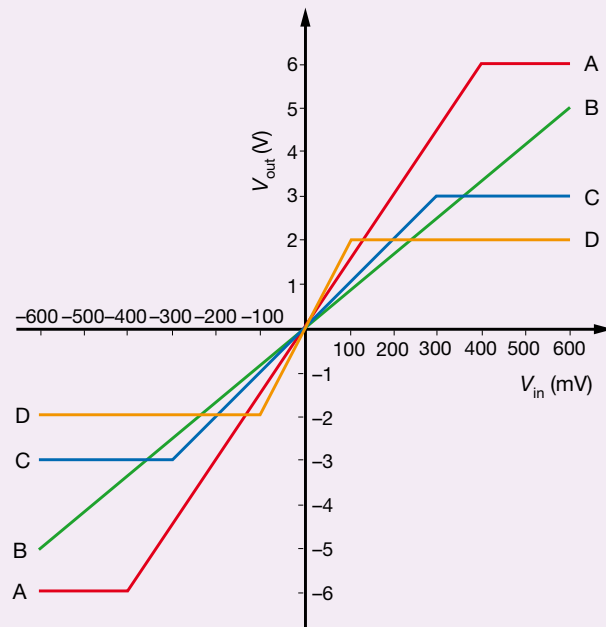


- 6 The following signal is fed into the input of the amplifier.



- Draw the output signal.
- Is the output signal a faithful amplified version of the input signal?
- Describe the output waveform in relation to the input signal.
- If the amplifier is used as part of an audio system, describe the quality of the sound that it would produce.

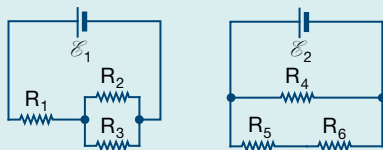
The following information applies to questions 7–10. The graph shows the characteristics of four different voltage amplifiers (A–D). The input voltage to these amplifiers is limited to between ± 600 mV.



- Which of the four amplifiers (A–D) has the highest linear amplification?
- Which of the four amplifiers (A–D) has the largest linear output range?
- Which of the four amplifiers (A–D) has linear amplification over the full range of input voltages?
- Which of the four amplifiers (A–D) has the smallest linear input range?

chapter review

- 1 In the following two circuits the batteries and resistors are identical. Assume the batteries are ideal (i.e. no internal resistance).



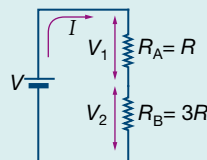
- Which resistor(s) has the highest current flowing through it?
 - Which resistor(s) has the lowest current flowing through it?
 - Which resistor(s) has the highest power dissipated through it?
 - Which battery is supplying the largest current?
- 2 You have four $4\text{ k}\Omega$ resistors. How would you arrange the four resistors to give a total effective resistance of:

- $16\text{ k}\Omega$?
- $10\text{ k}\Omega$?
- $1\text{ k}\Omega$?
- $5.33\text{ k}\Omega$?

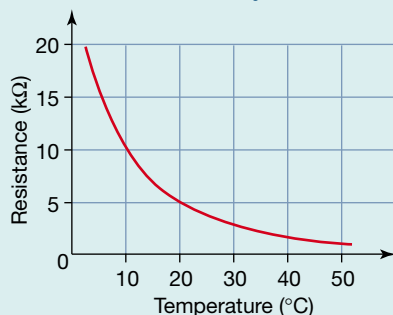
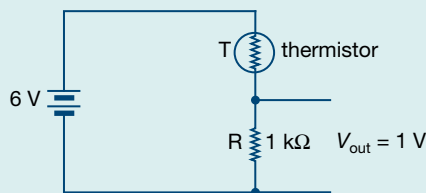
The following information applies to questions 3 and 4.

The circuit shows two resistors connected in series across a battery with a terminal voltage of V .

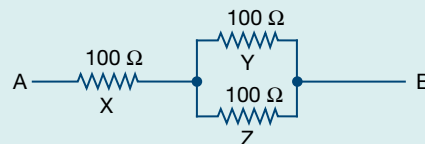
The resistance of R_B is three times that of R_A .



- Show mathematically that one-quarter of the battery's voltage is dropped across R_A .
- If $V = 12\text{ V}$ and $I = 200\text{ mA}$, determine:
 - the resistance values of R_A and R_B
 - the power dissipated in R_B .
- A thermistor is a semiconductor device whose resistance depends on the temperature. The graph shows the resistance versus temperature characteristic for a particular thermistor. Determine the temperature of the thermistor in the circuit if V_{out} is 1 V .

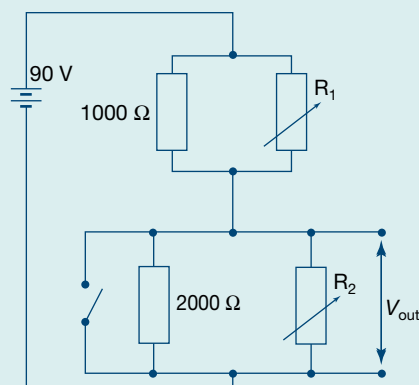


- 6 Three $100\text{ }\Omega$ resistors are connected as shown. The maximum power that can safely be dissipated in any one resistor is 25 W .

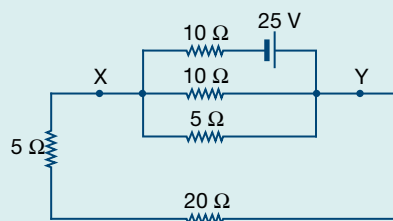


- What is the maximum potential difference that can be applied between points A and B?
 - What is the maximum power that can be dissipated in this circuit?
- 7 The circuit below combines variable resistors R_1 and R_2 with fixed resistors to make a complex voltage divider. Complete the table by determining the output voltage for each row.

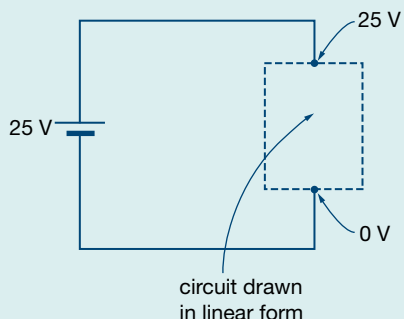
$R_1\text{ (}\Omega\text{)}$	$R_2\text{ (}\Omega\text{)}$	Switch	$V_{\text{out}}\text{ (V)}$
1000	2000	open	
2000	4000	open	
4000	2000	open	
8000	5000	closed	



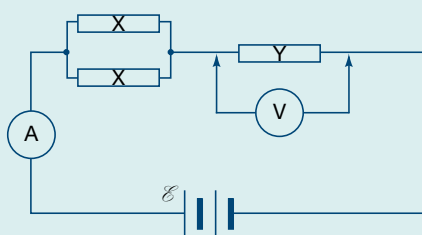
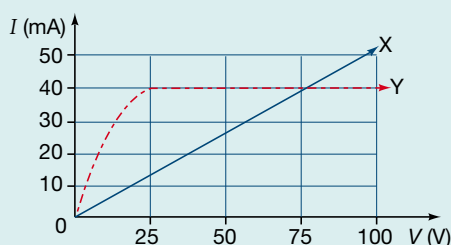
- Explain the meaning of the following statements about amplifier circuits.
 - At saturation the amplifier gain is not linear.
 - The amplifier circuit must be correctly biased before receiving an AC input signal.
- For the circuit shown, find:
 - the current in the $20\text{ }\Omega$ resistor
 - the potential difference between points X and Y.



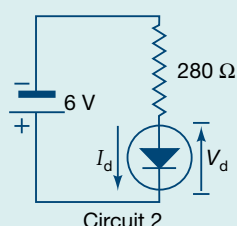
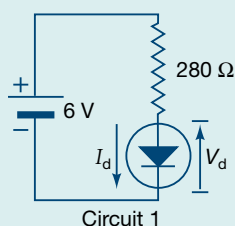
It may be useful to draw the circuit in a linear form, so that the voltage across it drops from top to bottom, as shown below.



- 10 Two electronic components, X and Y, are operating in the circuit shown. The voltage–current characteristics of each device are shown in the accompanying graph. The reading on the voltmeter is 60 V.

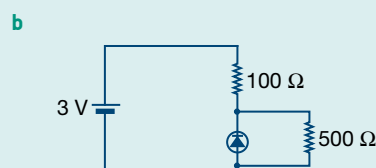
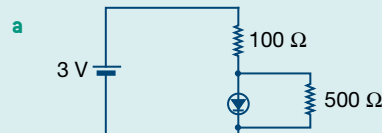


- Which device is non-ohmic?
 - What is the resistance of X?
 - What is the reading on the ammeter?
 - Determine the EMF of the battery.
 - Calculate the total power consumption in the circuit.
- 11 A silicon diode has a switch-on voltage of ~ 0.7 V and a thermal leakage current of ~ 100 nA. Determine the voltage across the diode (V_d) and the current (I_d) through the diode for:
- circuit 1
 - circuit 2, where the polarity of the battery has been reversed.

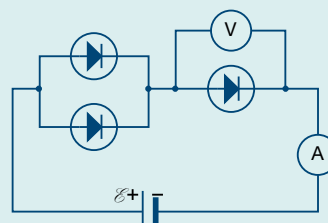
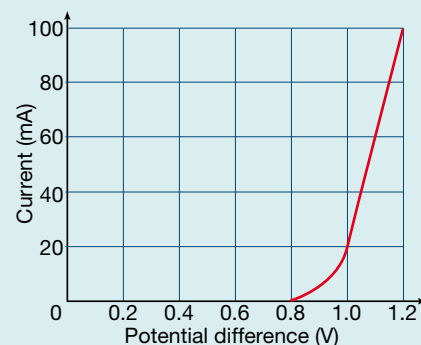


- 12 For the circuits shown in diagrams a and b, assume the diode has a switch-on voltage $V_s = 0.7$ V and a thermal leakage current $I_t = 1$ nA. Determine the current flowing through the:

- $100\ \Omega$ resistor
- $500\ \Omega$ resistor
- diode.



- 13 All the diodes in the circuit below are identical and have the voltage–current characteristics shown in the graph.

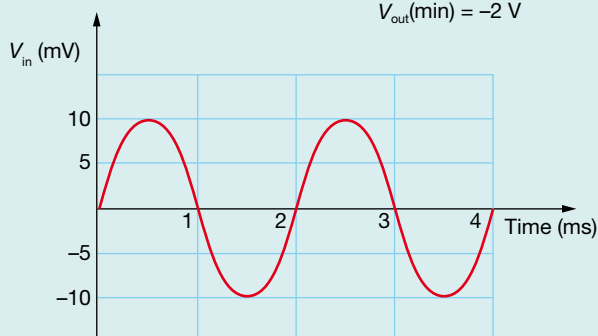
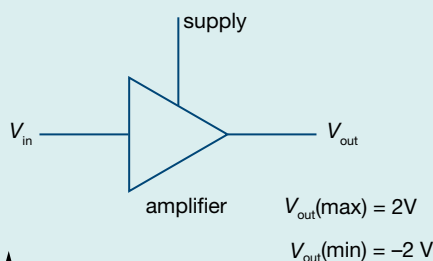


The ammeter reading is 52 mA.

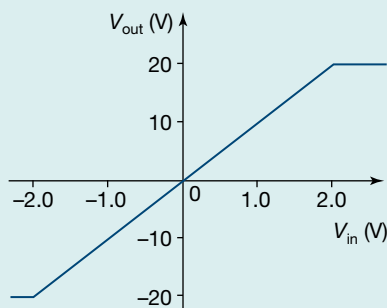
- What is the reading on the voltmeter?
- What is the EMF of the battery?

The following information applies to questions 14–16.

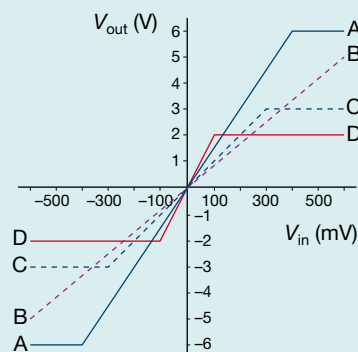
The amplifier circuit shown below has a voltage gain of -100 . Assume it is correctly biased so that an input of 0 V results in an output of 0 V . The small varying input signal shown below is then fed into the amplifier.



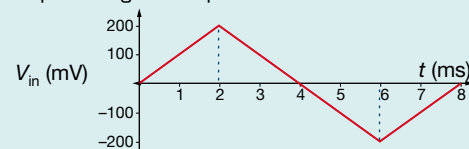
- 14 Quote the peak-to-peak voltage of the input signal shown.
- 15 Draw a graph of the resulting output voltage
- 16 The size of the input signal is now tripled but its frequency is unaltered. Draw the resulting output voltage graph.
- 17 The input–output characteristics of an amplifier are shown in the graph.



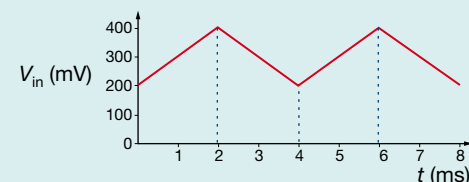
- a What type of amplifier has these characteristics?
 - b Calculate the voltage gain for that amplifier.
 - c What is the maximum peak-to-peak voltage that this amplifier can amplify without distortion?
- 18 The graph shows the characteristics of four different voltage amplifiers (A–D). The input voltage to these amplifiers is limited to between $\pm 600\text{ mV}$.



- a If the input voltage is as shown in the diagram, sketch the output voltage for amplifiers B and D.

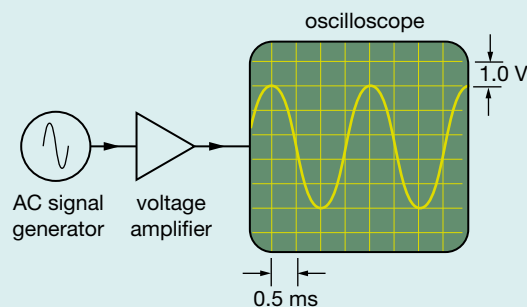


- b If the input voltage is as shown in the diagram, sketch the output voltage for amplifiers A and C.



The following information applies to questions 19 and 20.

A cathode ray oscilloscope is used to monitor the output of a linear amplifier. An AC signal generator is used to produce a sinusoidal input voltage with a peak-to-peak voltage of 50.0 mV .



- 19 a Is the amplifier operating within its limits? How can you determine this?
- b Analyse the gain of the amplifier.
- 20 a What is the frequency of the input signal?
- b What is the frequency of the output signal?



Introducing photonics



In this chapter, we study how the integration of electronics and photonics ideas has allowed the rapid development of a modern telecommunications system, including the Internet. We shall investigate (in a practical sense) the important ideas of converting a light signal to an electronic signal and vice versa, and transmitting information via a beam of light.



by the end of this chapter

you will have covered material from the study of photonics including:

- conversion of electrical signals into light signals
- comparing the information-carrying capacity of copper wires and optical fibres
- using electrical energy to generate light by using semiconductor materials
- detecting light through changes in the electrical properties of semiconductor materials
- using technical specifications for photonic devices to design and build simple photonic circuits
- transmitting information (voice or music) by a light beam.

CHAPTER 5

5.1

Photonics in telecommunications

Photonics is the science of using light to manipulate information and energy, two very important commodities in our modern world. Photonics is involved with all aspects of visible, ultraviolet and infrared radiation (i.e. its production, detection, transport, manipulation etc.). Photonics spans a vast array of optical phenomena where light is sometimes modelled as a stream of particles (called photons), each with a discrete amount (or quantum) of energy, and sometimes modelled as a continuous wave.

When light is modelled as a stream of photons, the energy of each discrete photon is very small (of the order of 10^{-19} J) and inversely related to its wavelength (or colour). It turns out that the energy of visible light photons is about the same as the energy needed to shift electrons about in crystal lattices or between the outer shells of many atoms. Hence, the photon description of light is important when light interacts with matter—as it does in photonic devices.

The discrete or 'grainy' photon nature of light only becomes apparent when we are looking at how light interacts with matter on an atomic scale or when the light has a very, very low intensity. In all other cases, we generally 'average out' the photon nature of light and model it as a continuous wave of a particular frequency and wavelength.

Light radiation is a small subset of a much more general phenomenon called electromagnetic radiation. Figure 5.1 shows the extent of the electromagnetic spectrum. Photonics deals with the optical spectrum, which is only a very small part of the electromagnetic spectrum. The optical spectrum covers visible light (i.e. electromagnetic radiation with a wavelength between 350 (violet) and 750 nm (deep red)), but also includes the non-visible near-infrared radiation (which extends above 750 nm) and near-ultraviolet radiation (which extends below 350 nm), as shown in Figure 5.1.

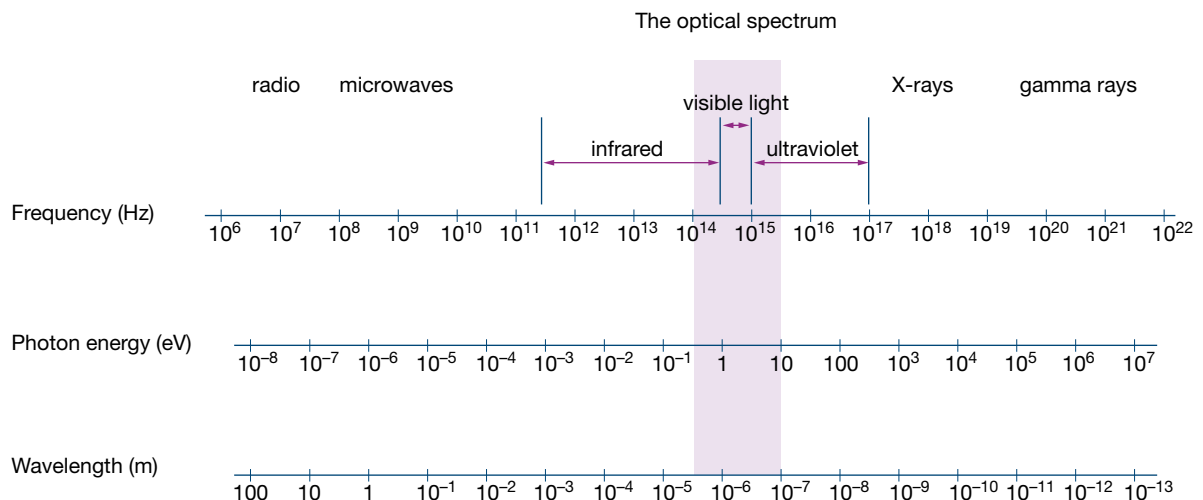


Figure 5.1 The electromagnetic spectrum. [Note that the electronvolt is a non-SI unit of energy commonly used in photonics: $1 \text{ eV} = 1.602 \times 10^{-19} \text{ J}$.]

These technologies are discussed in detail in Chapter 14.

The science of photonics has undergone a phenomenal development over recent years, and photonics-trained scientists and engineers are highly sought after by a diverse range of industries throughout the world. Photonics is the discipline responsible for the enormous growth in communications information-carrying capacity (or bandwidth) that has made the rapid expansion of the Internet possible. In a very real sense, photonics is the enabling technology for the current high bandwidth Internet developments such as media streaming.

In addition, photonics has expanded rapidly into a wide range of new technologies, in areas such as manufacturing and medicine. In this chapter, we will restrict our discussions to the electronics–photonics interface and the importance of photonics in telecommunications.

Information-carrying capacity and the skin effect

So how can the Internet infrastructure keep up with the ever-increasing rate of Internet traffic? If automobile traffic was increasing at the same rate, we could never build enough roads, freeways or traffic lights to keep up; and even if we could build all the infrastructure needed to meet the demand, our cities would soon be choked with concrete and asphalt. In the case of the Internet, infrastructure has been able to keep pace with the rapidly increasing traffic because of the huge information-carrying capacity of a photonics-based optical fibre network.

The information-carrying capacity of any telecommunications system is determined by the highest frequency that can be practically sent through the transmission medium. Around 40 years ago most telecommunications were transmitted as electrical signals, in the form of electromagnetic waves, travelling along copper cables. With copper wire, the maximum possible transmission frequency is limited by a physical phenomenon called the *skin effect*.

When low-frequency electrical signals propagate through a conductor (e.g. copper wire), the signal penetrates through the entire cross-section of the wire. Since the electrical resistance of the copper wire is inversely proportional to its cross-sectional area, low-frequency signals encounter only a small electrical resistance as they propagate through the wire. The copper wire has a relatively low resistance; so its transmission energy losses are small for low-frequency electrical signals. *Low-frequency signals propagate through the copper cable very efficiently with relatively low attenuation.*

The depth of penetration of the electrical signal (skin depth) decreases as the frequency of the signal increases. Thus, high-frequency signals only penetrate into the outer layer or ‘skin’ of the copper wire, and hence the effective cross-sectional area for high-frequency signals is very small

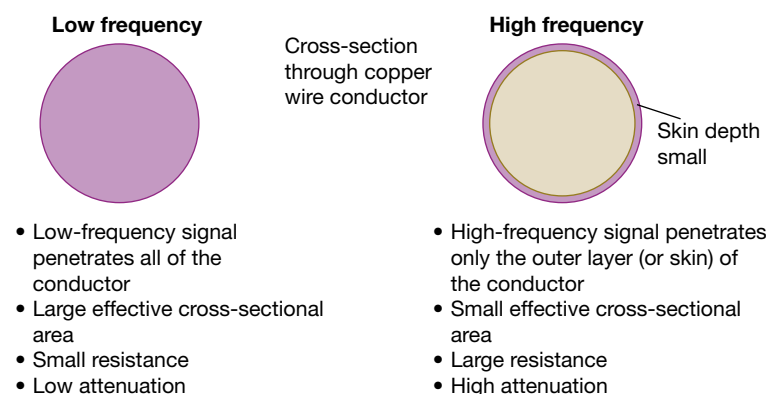


Figure 5.2 Skin depth for low- and high-frequency electrical signals propagating in a copper wire conductor.

Physics file

Attenuation is defined as a reduction in power level between two points as a signal travels through a medium. The loss of power is due to the interaction of the signal with the medium. We can think about the attenuation in terms of the loss of electrical power as an electrical signal travels down a copper wire, or the loss of optical power as a light signal travels down an optical fibre.

Electrical signals travelling down copper wires experience much more attenuation than light signals travelling down optical fibres. This means that the electrical signals need to be boosted more often (about every 5 km) than light signals (currently about every 50 km).

Physics file

The skin depth, δ , for electrical signals penetrating into a conductor is inversely proportional to the square root of the frequency (i.e. $\delta \propto \frac{1}{\sqrt{f}}$). Table 5.1 shows the skin depth as a function of frequency for a copper conductor. Even at frequencies as low as a few megahertz, the electrical signal penetrates only around a micron (i.e. one-thousandth of a millimetre) into the copper wire. The skin effect means that the maximum frequency that can be practically transmitted by a copper wire is a few tens of megahertz.

Table 5.1 Skin depth versus frequency for a copper conductor

Frequency (f) [Hz]	Skin depth (δ) [mm]
60	8.5
1×10^3	2.1
1×10^6	0.066
36×10^6	0.0012

Source: American Institute of Physics Handbook [McGraw-Hill] New York 1963.

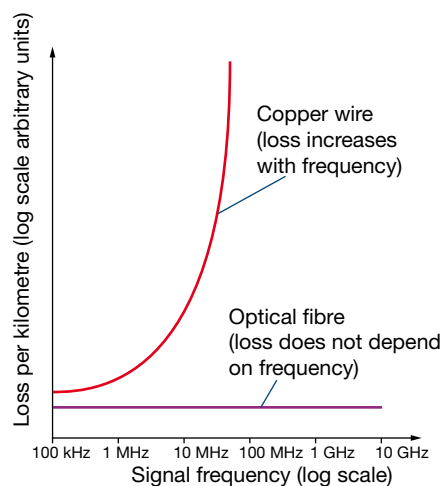


Figure 5.3 Graph of loss per kilometre versus frequency for copper wire and optical fibre.

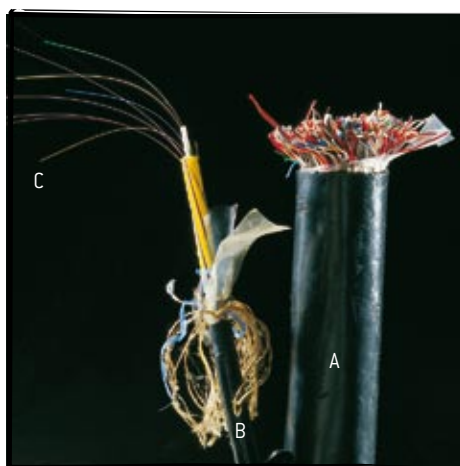


Figure 5.4 600 twisted-pair copper cable and optical fibre cable. A, copper telecommunication cable; B, fibre optic cable; C, single optical glass fibre.

(see Figure 5.2). High-frequency signals, therefore, encounter a large electrical resistance as they propagate through the wire, and the transmission energy losses are large. *High-frequency signals therefore propagate inefficiently with relatively high attenuation.*

Light signals channelled through optical fibres (by total internal reflection, as has been discussed in *Heinemann Physics 11*), do not suffer from the skin effect. This means that much higher frequencies (theoretically well beyond 10 GHz, i.e. 10^{10} Hz) can be transmitted through an optical fibre. Figure 5.3 compares the loss per kilometre as a function of frequency for a copper wire and an optical fibre. As the graph shows, optical fibre has a far superior information-carrying capacity than copper wire. Figure 5.4 shows a typical copper telecommunications cable. The cable consists of 600 pairs of twisted copper wire and has a total capacity of 600 voice conversations. It is bulky, heavy, difficult to handle and inefficient, considering the large quantities of raw materials used in its manufacture. Figure 5.4 also shows a typical fibre optic cable. This cable houses 30 single optical glass fibres, each one with a diameter of about 0.1 mm. Most of the bulk of the cable is used to mechanically support and protect the 30 'hair-like' glass fibres. Each one of the glass fibres has a capacity of 1000 voice conversations. Techniques currently being developed will increase the capacity of a single optical fibre to the equivalent of around 1 000 000 voice conversations!

Compared with twisted-pair copper telecommunication cables, optical fibre cables are lighter, more flexible and easier to handle, use less raw materials, are cheaper to produce and have a much higher information-carrying capacity. As well, the number of individual optical fibres that can be housed in a cable can easily be increased to many hundreds or perhaps even thousands if ever required. Optical fibres are very cost effective, and they have an enormous information-carrying capacity. Networks of optical fibres are the only telecommunication infrastructure system that could possibly support the current and future rapid expansion of the Internet.

Interestingly enough, high-frequency electrical signals can be carried over relatively short distances (up to hundreds of metres) by transmission lines (e.g. a central solid copper wire surrounded by a cylindrical copper conductor) or waveguides (e.g. a hollow cylindrical copper conductor), in which the signal is transmitted by the electromagnetic field in the free space inside the cylindrical conductor.

The simplest example of a high-frequency transmission line is the coaxial cable connecting your TV aerial to your television set.

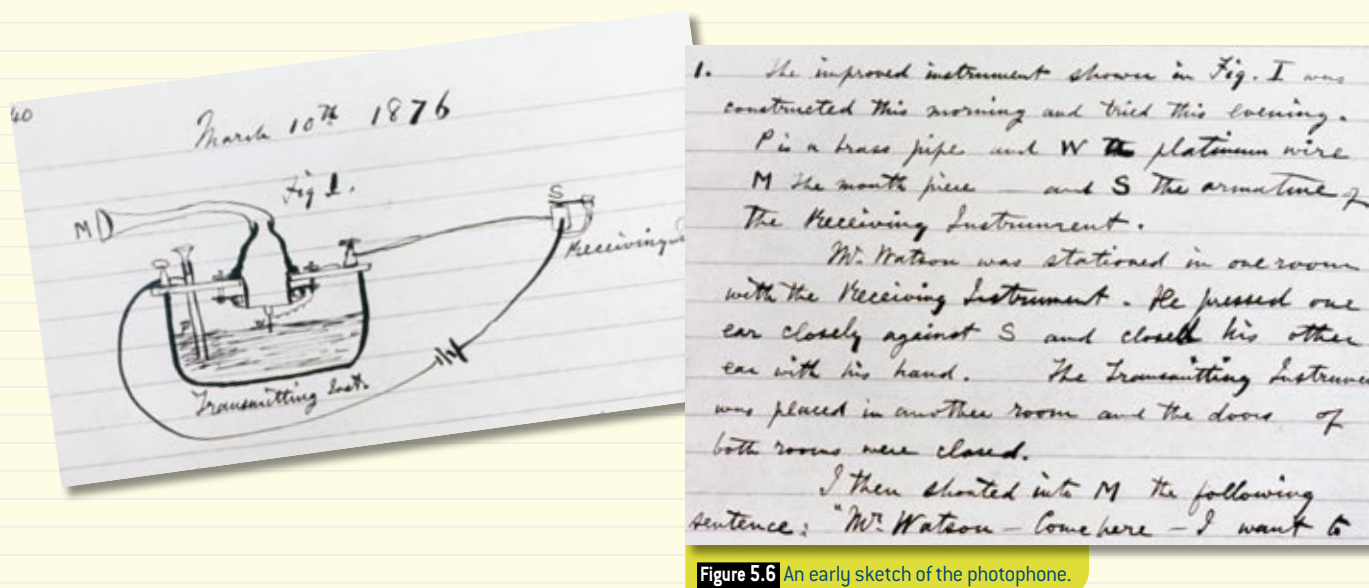
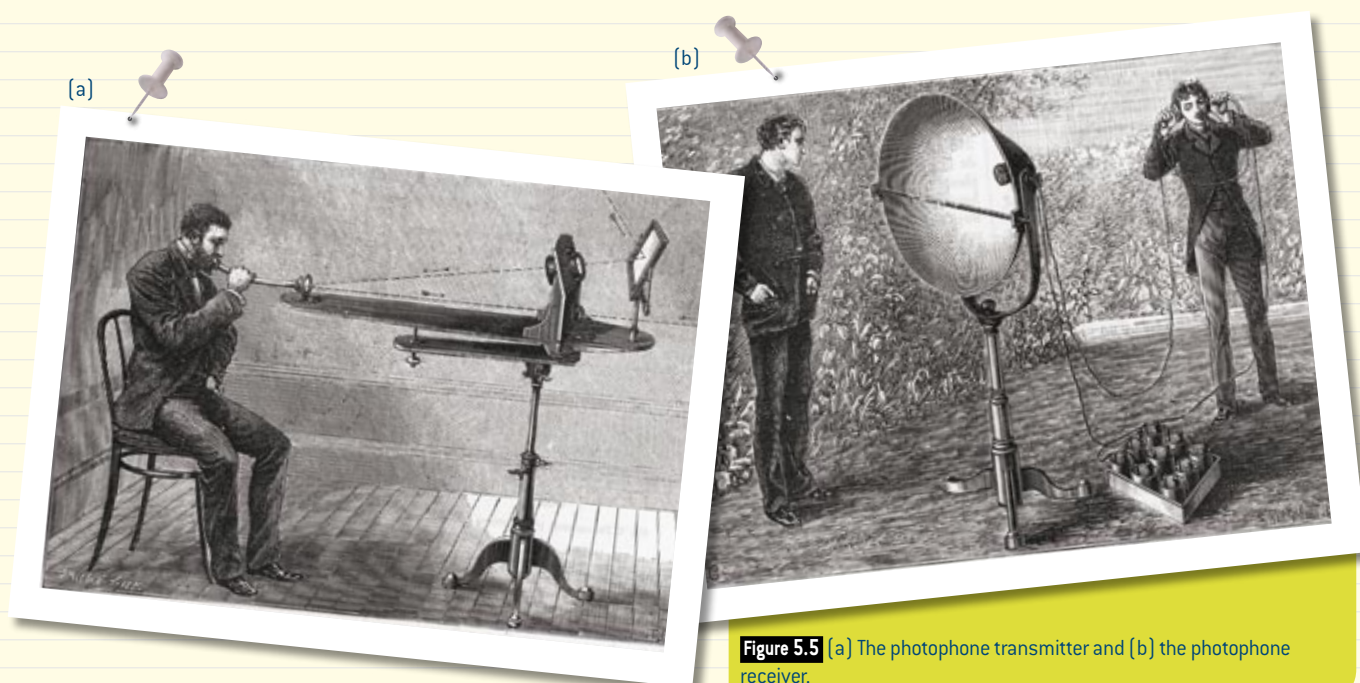
From the photophone to the Internet

In 1876, Alexander Graham Bell invented the electric telephone. The electric telephone had an immediate and long-ranging influence on society. It is unquestionably one of the great inventions of the post-Industrial Revolution era. In 1880, just 4 years after the invention of the telephone, Bell also invented the 'photophone'. The photophone was also a telecommunication device that could transmit a voice signal from one point to another; but instead of using electricity moving in a copper wire, the photophone used light travelling in air as the transmission medium. A more detailed description of the photophone can be found in the following Physics in action. Ahead of its time, this invention was essentially ignored and the idea of using light in modern telecommunication systems lay dormant for almost 100 years.

The photophone

Alexander Graham Bell invented the photophone in 1880, just 4 years after he invented the electric telephone. The photophone was a telecommunications device that could send a voice signal with relatively good fidelity over a distance of about 200 m by using a light beam as the information carrier. Figures 5.5 and 5.6 show early sketches and Bell's original hand-drawn diagrams of the photophone. Figure 5.7 details the main components of the photophone.

Sunlight (or lamplight) is reflected by a mirror (on the transmitter) through a lens, which focuses the light onto a thin, flexible mica sheet (diaphragm). The diaphragm reflects the incident light beam through a second lens and then onto a parabolic collector (on the receiver). The diaphragm is connected to a mouthpiece via a tube that concentrates the sound energy. Sound



pressure (from the mouthpiece) vibrates the diaphragm, which in turn slightly deflects the beam of reflected light from its original path. Thus, the intensity of light illuminating the parabolic collector varies depending on the amplitude of the voice signal. This is known as intensity (or amplitude) modulation: the intensity of the light beam (which carries the information) is proportional to the amplitude of the voice signal (the information).

The parabolic collecting mirror focuses the modulated light beam onto a selenium cell that is mounted at the mirror's focal point. Some years earlier, it had been discovered that selenium had a very interesting property: its resistance depended on the intensity of the light illuminating it. The selenium cell is connected in series with a battery and earpiece. The modulated light intensity on the selenium cell varies the current through the earpiece circuit, and generates a copy of the original sound at the receiver.

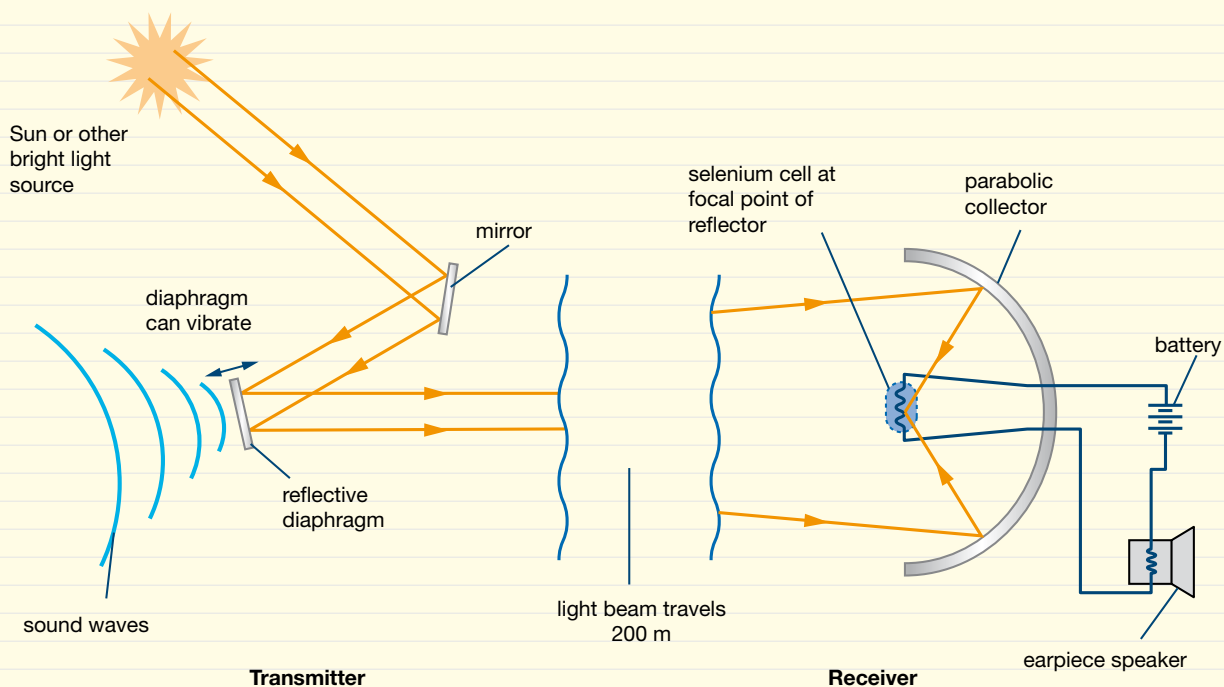


Figure 5.7 The main components of the photophone.

The photophone was not an early commercial success for two reasons. First, the light sources available for the photophone (Sun or lanterns) were not reliable or bright enough. The original photophone only had a range of about 200 m even with the most intense light source available at the time (the Sun). It was not until the 1960s, when the first lasers were built, that a truly stable and bright light source was available for photonics-based telecommunications. Lasers provided a bright and easily controlled light source, the optical energy of which could be directed into a narrow, well-defined path from the transmitter to the receiver. Second, the 'line-of-sight' air path used in the photophone was inherently unstable. Fog, rain or even stray animals could easily interrupt the light beam that carried the telecommunication signal. It was only in the 1970s, when ultratransparent glass optical fibres became available, that a stable, low-loss medium could be used for channelling the light in photonics-based telecommunication systems.

Normal glass is not very transparent. Take a sheet of window glass and look at it side on: very little light gets through even over distances as small as 10 cm. Because of its high level of light scattering and absorption, normal glass is totally unsuitable for transmitting light over long distances, even if the light is channelled in an optical fibre by total internal reflection! By the 1960s, improved glass fibre technology meant that light absorption and scattering was substantially reduced. Even so, the best transmission that could be achieved with optical fibres of that era was only 1% over a distance of 20 m. In other words, only 1% of the channelled light intensity entering one end of the fibre was still being channelled 20 m further on. The loss of 99% was due to scattering and absorption of light by impurities and imperfections in the glass. During the 1970s, many advances were made in the production of high-purity silica glass optical fibres. At the start of that decade, glass fibres were available that could transmit 1% over 1 km; by the end of the decade, optical fibres were being manufactured that could transmit 95.5% over 1 km (which is close to the theoretical limit). To put this in perspective, if seawater was as transparent as these modern optical fibres, we could easily see to the bottom of the Mariana Trench—the deepest point on Earth, approximately 11 km below the surface of the Pacific Ocean. Given the low transmission loss of optical fibres and their high information-carrying capacity, optical fibres were the obvious choice for the backbone infrastructure of the newly developing Internet.

The Internet is a distributed system of interconnected nodes usually linked by optical fibres. The information stream that is to be transmitted is divided into a large number of information packets. Each packet has a header (which identifies the type of data, its source and its destination) and an information segment (which holds a given amount of the information stream). Each packet can travel through the Internet by using one of various combinations of nodes to get to its final destination (where all the packets are reassembled into the original information stream). This sort of transmission mode is called *asynchronous transfer*, as each information packet can traverse the Internet by whichever path happens to be the most efficient at that particular instant.

Although the Internet (now running on the optical fibre network) is highly successful and will continue to grow at a rapid rate, other small, laser-based 'line-of-sight' telecommunication systems are currently being developed. These systems, which are being manufactured in Australia and overseas, are a modern extension of Bell's original photophone. In the central business district (CBD) of many large cities, laser beams transfer large amounts of information from the top of one building to the next. These systems use powerful infrared laser beams that can penetrate fog and rain, at least over the short distances (up to 1 km) between buildings. These systems provide a short-range, dedicated, cost-effective alternative to the Internet, and allow companies to share and swap information without having to go through the Internet. An example of a modern laser 'line-of-sight' telecommunication system is shown in Figure 5.8.



Figure 5.8 Laser 'line-of-sight' telecommunication system.



5.1 summary

Photonics in telecommunications

- Light can be modelled as a continuous wave or as a stream of particles (photons), each with a discrete amount (quantum) of energy.
- Optical radiation covers the near-ultraviolet, visible and near-infrared regions of the electromagnetic spectrum.
- The rapid growth of the Internet is due to the very high information-carrying capacity of optical fibres.
- With solid copper wires, high-frequency electrical signals are highly attenuated because of the skin effect.
- Light signals in optical fibres are unaffected by the skin effect.
- The Internet is a distributed communications system, usually linked by optical fibres.



5.1 questions

Photonics in telecommunications

- Photonics can be considered to be the science of:
 - using light in telecommunications
 - using light to manipulate information and energy
 - understanding photography
 - studying small particles of matter.
- The 'optical spectrum' includes the following components of the electromagnetic spectrum:
 - visible, TV, near ultraviolet
 - near-infrared, visible, near-ultraviolet
 - near-ultraviolet, X-rays, visible
 - visible, near-infrared, microwaves.
- The rapid growth of the Internet has been possible because optical fibres:
 - have a very high information-carrying capacity
 - are much smaller than copper cables
 - do not use copper, which is a very expensive raw material
 - have a much longer lifetime than copper cables.
- The skin effect means that:
 - low frequencies in copper cables and low frequencies in optical cables have high transmission losses
 - high frequencies in copper cables and high frequencies in optical cables have low transmission losses
 - high frequencies in copper cables and low frequencies in optical cables have low transmission losses
 - low frequencies in copper cables and high frequencies in optical cables have low transmission losses.
- The skin depth of an electrical signal:
 - increases with increasing frequency in copper wire
 - increases with increasing frequency in optical fibre
 - decreases with increasing frequency in optical fibre
 - decreases with increasing frequency in copper wire.
- The photophone was invented:
 - just a few years after the electric telephone
 - just a few years before the electric telephone
 - approximately 50 years after the electric telephone
 - approximately 100 years after the electric telephone.
- Modern silica optical fibres used in telecommunications can transmit:
 - 1% of the initial signal over 1 km
 - 1% of the initial signal over 20 m
 - 80% of the initial signal over 1 km
 - 95.5% of the initial signal over 1 km
 - 95.5% of the initial signal over 200 km.
- The Internet:
 - uses a continuous, uninterrupted bit stream to transmit from transmitter to receiver
 - uses two types of information packets—one for the address and one for the information
 - is a distributed communications system
 - uses optical fibres, which can be easily moved from one node to the next.
- The photophone had the following problems limiting its development:
 - unreliable bright light source and unstable transmission medium
 - insufficient funding to develop the prototype
 - modulation system unsuitable for audio transmission
 - light detection system unsuitable for audio transmission.

5.2 Optical transducers

In modern telecommunication systems, the important interface between electronics and photonics largely relies on the development of *optical transducers*: devices that convert light energy (which often can contain information) into electrical energy and vice versa. This conversion of one form of energy to another is important because although information can be transferred from one point to another faster if it is encoded into light, it can be manipulated much more easily in electronic form. This means that we can interpret the information stored in a light beam much more easily and accurately once it has been converted to an electrical signal.

Let's look at an example to show how an optical transducer can help us correctly interpret information stored in a light signal. Switch on an incandescent light globe (connected to the AC mains power supply) and what do we see? Our senses would have us believe that we are observing a *constant* light intensity (as shown in Figure 5.9)!

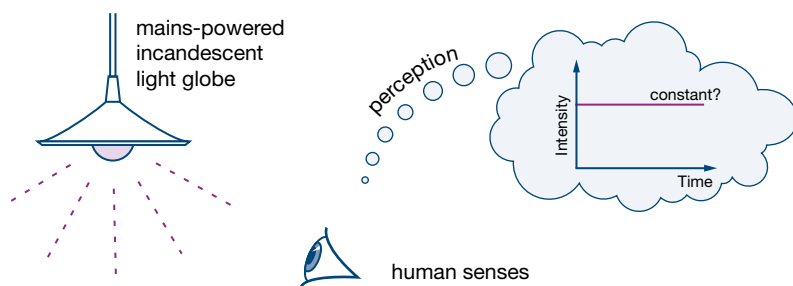


Figure 5.9 What do we see when we 'look' at a light globe?

In reality, however, the intensity of the light globe is not constant but varies periodically with a frequency of 100 Hz. With a mains-operated AC light globe, the sinusoidal alternating current (i) flowing through the filament has one forward and one backward peak in each complete cycle (see Figure 5.10a) and has a frequency of 50 Hz. Each peak in current (both forward and backward) generates enough heat in the filament to make it glow 'white' hot. We therefore have two peaks in the light intensity of the light globe for each alternating current (AC) cycle. This is why the light intensity (I) has a frequency of 100 Hz (Figure 5.10b). Although the calculation is beyond the scope of this study, it can be shown that the light intensity, like the current, is a sinusoidal oscillation. Our senses (in this case, our human eyes) are *unable* to detect this 100 Hz oscillation in light intensity and simply average the true signal, as shown by the dashed line of Figure 5.10b.

We can use a photonics sensor (in this case, a simple solar cell, which is normally used to change light energy into electrical energy) to convert some of the light emitted by the globe into an electrical signal. We can then manipulate the electrical signal into a more easily interpreted signal by using an electronic measuring instrument (in this case, the plot from a datalogger, as shown in Figure 5.11). The datalogger digitally takes a 'snapshot' of the periodic, time-varying signal. The snapshot is then displayed on a screen in a way that allows us to easily observe the true periodic nature of the light intensity from the light globe. Note that the solar cell converts the light intensity into an electrical signal in a *non-linear* manner, and hence the voltage displayed on the oscilloscope is a slightly distorted 100 Hz sinusoidal signal.

Physics file

A sinusoidal oscillation $y(t)$ has the following mathematical form:

$$y(t) = Y_0 \sin(2\pi ft)$$

where Y_0 = the amplitude of the oscillation

f = its frequency

t = time.

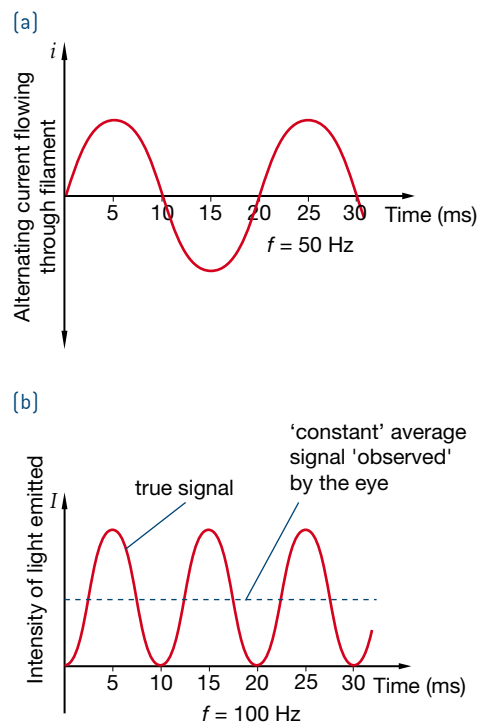


Figure 5.10 (a) Sinusoidal current through a light globe. (b) Output light intensity and average light intensity.

Figure 5.11 Measuring light intensity with a solar cell and a datalogger.



Optical detectors

Photoconductive detectors: Light-dependent resistors

One of the simplest optical sensors (i.e. optical to electrical transducers) is the photoresistor or light-dependent resistor (LDR). The LDR is made of a semiconductor material (often cadmium sulfide), whose resistance changes when it is illuminated by light of different intensities. In an LDR, a channel of cadmium sulfide (CdS) is formed into the shape of a long zig-zag line (see Figure 5.12a). A thin layer of transparent plastic material covers the semiconductor material so that light can be absorbed into the semiconductor. As the light intensity increases, the semiconductor's resistance decreases. A typical LDR might have a resistance as low as tens or hundreds of ohms in direct sunlight, and as high as several million ohms in total darkness. The circuit symbol for an LDR is shown in Figure 5.12b.

The curve in Figure 5.13 shows how resistance varies with illumination for a typical LDR. The curve is plotted on a log-log scale and so the graph indicates that the resistance of the sensor does *not* vary linearly with illumination. The LDR is very sensitive since it exhibits a relatively large change in resistance



Figure 5.12 (a) Two different LDRs; (b) circuit symbol for an LDR.

Physics file

Response times given in this book are for the optical sensors themselves, and can be longer in certain detector circuit configurations.

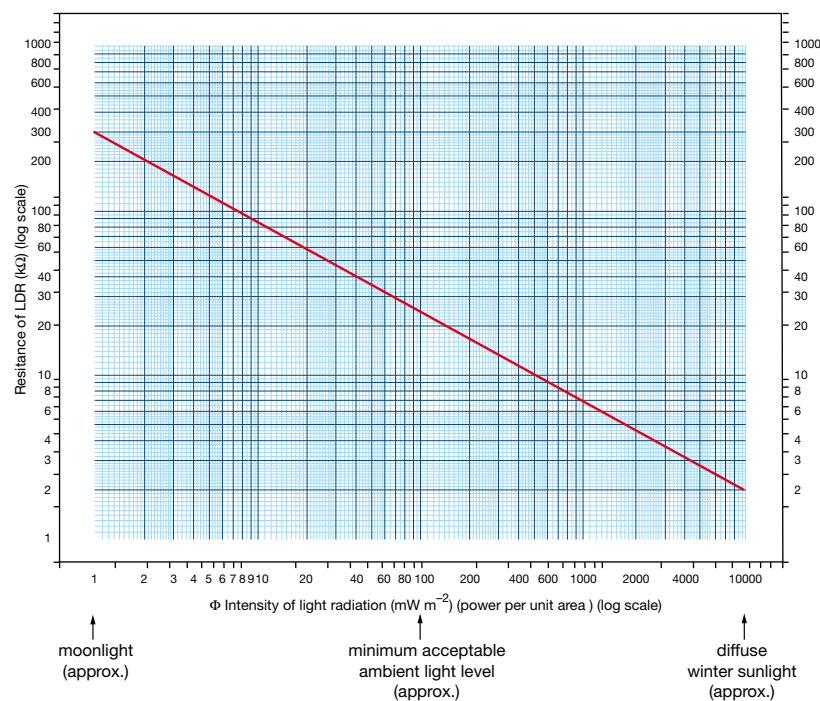


Figure 5.13 Resistance versus illumination for a typical LDR.

for a small change in illumination (especially at low light intensities). By connecting the LDR in series with another fixed resistor, we can make a voltage divider circuit, where the output voltage is a measure of the light intensity illuminating the LDR. Remember that LDRs are only suitable for measuring relatively slow changes in light intensity, since their response time is relatively slow. Generally, LDRs are suitable for detecting light signals that vary down to a time scale of milliseconds.

Worked example 5.2A

The circuit shown is used to switch on (off) a street lamp when the light intensity falls below (rises above) a certain trigger level. V_{cc} is the voltage of the DC power supply for the circuit. Assume that the transistor is saturated (very strongly switched on) when $V_{be} > 0.7$ V, and off when $V_{be} < 0.7$ V.

The street lamp should be switched on when the light intensity falls below 100 mW m^{-2} . The LDR has the same characteristics as shown in Figure 5.13. Assume that a $V_o [V_{out}] > 5$ V [< 1 V] switches the lamp on (off). Determine the values of R_2 and R_c , given that the transistor has a current gain $I_c/I_b = 100$.

Hint: It might be useful to revise transistor amplification and biasing in Chapter 4.

Solution

If the base current, I_b , is small, then the base–emitter voltage, V_{be} , is essentially determined by the LDR/ R_2 voltage divider.

From the curve of LDR characteristics, when the light intensity is 100 mW m^{-2} , the resistance of the LDR is $25 \text{ k}\Omega$. If the light intensity falls below this level, R_{LDR} increases and V_{be} decreases. Recall that V_{be} is determined by the LDR/ R_2 voltage divider:

$$\text{i.e. } V_{be} = V_{cc} \frac{R_2}{R_2 + R_{LDR}}$$

If V_{be} decreases (i.e. light intensity falls), the transistor will tend to turn off and the current [I_c] through R_c will fall. Hence V_o will increase to be ≥ 5 V and the street lamp will turn on.

So we need to ensure that when $R_{LDR} = 25 \text{ k}\Omega$, the transistor is just turning off (i.e. V_{be} is a little less than 0.7 V). To be safe we need to choose a value of R_2 so that when R_{LDR} is $25 \text{ k}\Omega$ the voltage at the base of the transistor $V_{be} = 0.65$ V.

Assuming I_b to be small compared with the current flowing through the R_{LDR}/R_2 voltage divider, we have:

$$V_{be} = V_{cc} \frac{R_2}{R_2 + R_{LDR}}$$

Given that $V_{be} = 0.65$ V when $R_{LDR} = 25 \text{ k}\Omega$, we can calculate R_2 :

$$0.65 = 6 \times \frac{R_2}{R_2 + 25 \times 10^3}$$

$$0.65R_2 + 0.65 \times 25 \times 10^3 = 6R_2$$

$$16.25 \times 10^3 = 5.35 R_2$$

$$R_2 = 3.04 \text{ k}\Omega$$

So, if $R_{LDR} > 25 \text{ k}\Omega$

light level $<$ trigger level

$V_{be} < 0.7$ V

transistor off

$I_c \approx 0$ mA

$V_{R_c} \approx 0$ V

$V_{ce} > 5$ V

$V_o > 5$ V

Street lamp on (as required)

but if $R_{LDR} < 25 \text{ k}\Omega$

light level $>$ trigger level

$V_{be} > 0.7$ V

transistor on (saturated)

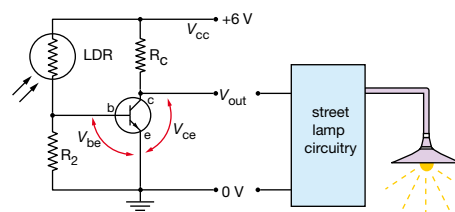
I_c is maximum

$V_{R_c} \approx 6$ V

$V_{ce} \approx 0$ V

$V_o \approx 0$ V

Street lamp off (as required)



Physics file

Semiconductors can also be used to make temperature sensors. As the temperature increases, thermal vibrations in the semiconductor lattice become more intense and more lattice bonds are broken. This results in an increasing number of mobile charge carriers.

Thus the conductivity of the semiconductor increases as the temperature rises. Such devices are called temperature-dependent resistors (TDRs) or thermistors.

Physics file

An LDR is made of an undoped semiconductor material (a compound of cadmium and sulfur, cadmium sulfide, is often used). As we discovered earlier, the atoms in a semiconductor are strongly bonded into a lattice. The resistance of semiconductors at room temperature is usually very high as there are very few mobile charge carriers in the lattice. These mobile charge carriers are electrons and holes created when thermal vibration occasionally breaks one of the lattice bonds.

If the energy of a photon is greater than the energy required to break a lattice bond in the semiconductor material, then the photon can be absorbed, creating an additional mobile electron and hole. This process will increase the number of number of mobile charge carriers and therefore decrease the resistance of the semiconductor material.

Physics file

Diffusion is when mobile electrons move from the n-doped region (high electron concentration) into the p-doped region (low electron concentration) at a semiconductor junction. Holes diffuse in the opposite direction. This is a similar process to the spreading of a strong odour from a region of high concentration to surrounding regions of low concentration.

Recombination is when holes and electrons recombine and release energy.

When the transistor is on (i.e. finite I_b) $R_{LDR} < 25 \text{ k}\Omega$; hence the current flowing through the LDR is:

$$\begin{aligned} I_{LDR} &= \frac{V_{LDR}}{R_{LDR}} \\ &> \frac{6 - 0.7}{25 \times 10^3} \\ &> \frac{5.3}{25 \times 10^3} \\ &> 0.2 \text{ mA} \end{aligned}$$

As we require $I_{LDR} \geq 10I_b$, choose $I_b > 0.02 \text{ mA}$. Since the transistor has a current gain of 100, this means $I_c < 2 \text{ mA}$. So our value of R_c is given by:

$$\begin{aligned} R_c &= \frac{V_{cc} - V_{ce}}{I_c} \\ &= \frac{6 - 0}{2 \times 10^{-3}} \\ &= 3 \text{ k}\Omega \end{aligned}$$

Junction detectors: Photodiodes

As we discovered in Chapter 4, a pn junction is formed when a semiconductor has adjacent p-doped and n-doped regions. Because of diffusion and recombination, a region that is depleted of mobile charge carriers (the depletion region) is created at the junction. An electrical potential barrier (V_b) and an electric field exist across the depletion region. The electronic signal diode discussed in Chapter 4 was made from a pn junction of doped silicon semiconductor material. Recall that when this type of diode is reverse biased, the externally applied voltage increases the effective potential barrier and virtually no charge carriers can diffuse across the junction (see Figure 4.16, page 133). Nevertheless, thermal vibrations of the silicon atoms will occasionally break a bond in the depletion layer. This creates mobile charge carriers (i.e. electrons that are strongly attracted to the p-side of the junction and holes that are strongly attracted to the n-side) that form a small reverse-biased current (called the thermal leakage current, I_l).

In a photodiode, the pn junction is constructed in such a way that its p-region is very thin and very close to the surface of the photodiode (see Figure 5.15a). This means that the semiconductor junction is located near the surface, and that there is a high probability that any light illuminating the surface can penetrate down and be absorbed in the depletion region. Figure 5.15b shows a photograph of photodiode. The square piece of semiconductor material, which is the 'active area' of the photodiode, can be clearly seen. A thin wire from the front surface of the semiconductor is electrically connected to one terminal, while the back surface is similarly connected to the metal case, which itself is connected to the other terminal. Figure 5.15c shows the circuit symbol for the photodiode.

We will see in Chapter 6 that when light is modelled as a stream of photons, the energy of each discrete photon is related to the frequency or wavelength of the light:

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

where f = frequency of the light (Hz)

λ = wavelength of the light (m)

c = speed of light in a vacuum ($2.998 \times 10^8 \text{ m s}^{-1}$)

h = Planck's constant ($6.626 \times 10^{-34} \text{ J s}$).

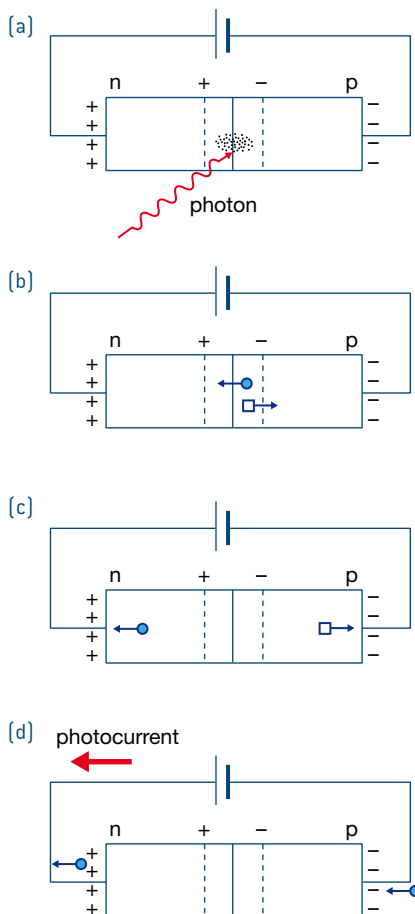


Figure 5.14 Semiconductor junction with photon absorbed and hole-electron pair being swept across depletion region.

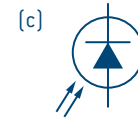
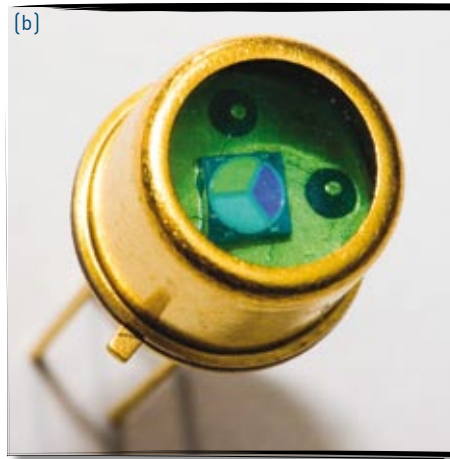
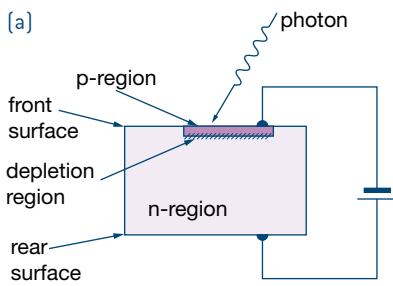


Figure 5.15 (a) A diagram of a cross-section through a photodiode; (b) a photograph of a photodiode; and (c) a circuit symbol for a photodiode.

If a photon (visible or near-infrared) is absorbed in the depletion region of a reverse-biased silicon photodiode, it may have the right amount of energy (E_p) to break one of the silicon lattice bonds in the semiconductor material, and hence create a free electron/hole pair (in much the same way that thermal energy creates the occasional free electron/hole pair). Under the influence of the large reverse-biased voltage, the mobile electron will be attracted towards the more positive p-doped region (i.e. the electron will be swept across into the p-doped region). The hole can also be considered to be a mobile charge carrier and is attracted to the more negative n-doped region (i.e. the hole is swept across into the n-doped region). The absorption of the photon thus brings about a movement of charge across the depletion region, as shown in Figure 5.14. The greater the intensity of the absorbed light (i.e. greater number of photons), the greater the number of charge carriers swept across the junction. The movement of holes and electrons through the diode generates a significant reverse-biased current, called the *photocurrent*,

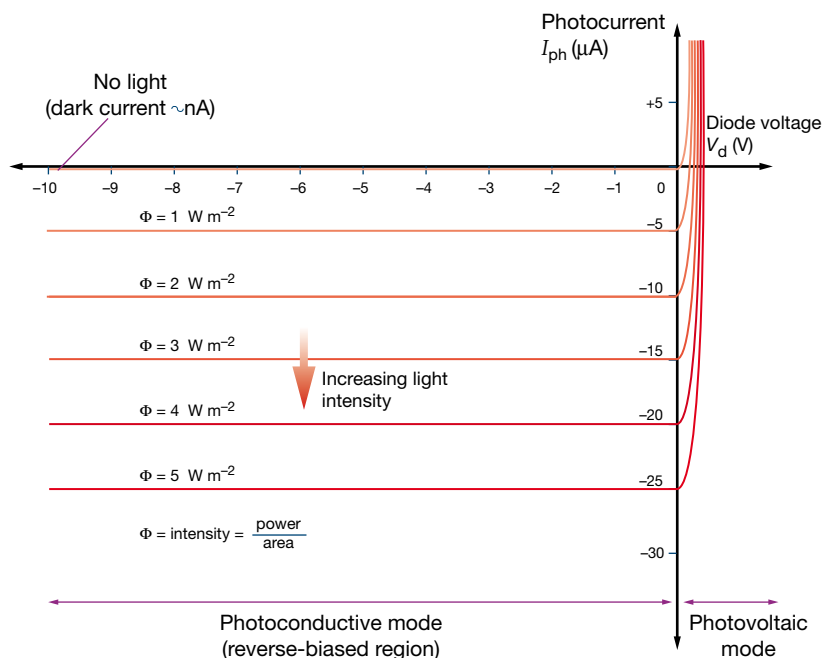


Figure 5.16 I - V characteristics of a typical photodiode.

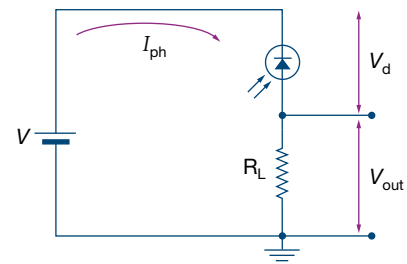
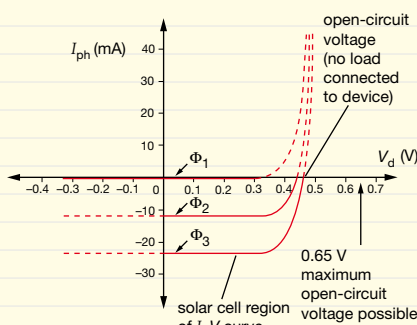


Figure 5.17 Simple photodiode detector circuit.

Physics file

Even if a photodiode is not reverse biased, any light absorbed in the depletion layer still gives rise to a separation of mobile charge carriers. A measurable voltage (or EMF) will be generated across the unbiased terminals of an illuminated photodiode, although this voltage is not linearly proportional to the light intensity (see 'open-circuit' region of the I - V characteristics in Figure 5.18). A voltmeter or oscilloscope connected across the photodiode will give an indication of the light intensity illuminating the depletion layer. For a silicon-based photodiode, the maximum potential difference that can be generated across the open-circuit terminals (i.e. there is essentially no connection between the terminals so that no current is being drawn) is approximately 0.65 V, even with very high-intensity illuminating light. An unbiased photodiode is said to operate in photovoltaic mode, as a voltage across the semiconductor junction is generated because of the illuminating light intensity.

A solar cell is essentially an unbiased photodiode with a very large and efficient light-absorbing semiconductor junction. When the solar cell is illuminated by an intense light (i.e. the Sun) it can be used as a source of EMF (i.e. like a battery). The maximum terminal voltage of a single silicon solar cell is about 0.65 V, although several cells can be connected in series to produce a higher terminal voltage (see Figure 5.19.)



Φ_1 = darkness (i.e. zero intensity)

$\Phi_2 = \frac{1}{2} \Phi_3$

Φ_3 = sunny day (600 W m^{-2})

Figure 5.18 I - V characteristics of a solar cell.

that is directly proportional to the illuminating light intensity. The current-voltage characteristics for a photodiode are shown in Figure 5.16 for a number of light intensities. Note that the 'dark current' curve is very similar to what we would get for a normal diode. As the light intensity increases, we get the additional contribution from the light-generated photocurrent. A reverse-biased photodiode operates in what is called *photoconductive mode*, since the conduction of the semiconductor junction varies with the illuminating light intensity.

If the reverse-biased voltage is relatively large (i.e. several volts) the potential difference across the depletion region will be large and the electrons and holes created by the absorption of photons will be swept across the junction very quickly. This means that the reverse-biased photodiode will have a very fast response time (much faster than an LDR) and is suitable for detecting light signals that vary down to a time scale of a fraction of a microsecond. The simple circuit shown in Figure 5.17 can be used to measure light intensity by using a photodiode.



Figure 5.19 [a] A single solar cell. [b] A bank of solar cells used to power an entire house.

Worked example 5.2B

The photodiode shown in the circuit has the same I – V characteristics as that shown in Figure 5.16. The ambient light intensity illuminating the detector is 3 W m^{-2} , and the detector's active area is $2 \text{ mm} \times 2 \text{ mm}$.

- Determine the value of R_L that ensures the detector is reverse biased by 4 V under this ambient illumination.
- Determine the radiation power absorbed by the active area of the detector under this ambient illumination.

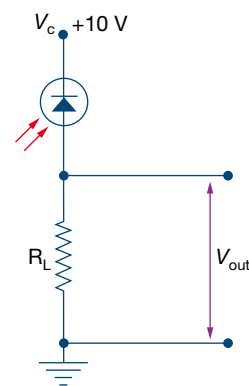
Solution

- From the curve of photodiode characteristics, an intensity of 3 W m^{-2} gives a photocurrent of $\approx 15 \mu\text{A}$ (for the reverse-biased photodiode).

To achieve a reverse bias of 4 V across the detector, 6 V must appear across R_L . Hence:

$$R_L = \frac{6}{15 \times 10^{-6}} \\ = 400 \text{ k}\Omega$$

- Intensity (ϕ) = $\frac{\text{power}}{\text{area}}$
 \therefore Power = intensity \times area, where area = $2 \text{ mm} \times 2 \text{ mm} = 4 \times 10^{-6} \text{ m}^2$
 $= 3 \text{ W m}^{-2} \times 4 \times 10^{-6} \text{ m}^2$
 $= 12 \times 10^{-6} \text{ W}$
 $= 12 \mu\text{W}$



PRACTICAL ACTIVITY 21

Solar cell's response to light

Physics in action

Solar-powered traffic monitors

Strange solar-powered stations (like the one shown in Figure 5.20) are appearing on many new motorways throughout Victoria. These unusual structures are part of a new road traffic-monitoring system that has been developed by VicRoads. Buried beneath each lane of the motorway, inductance coils detect passing cars. The station continuously monitors traffic flow, including vehicle speed, length and volume. Wires from the coils are connected to a microprocessor in the control box (labelled A in Figure 5.20) at the base of the station. Information about traffic flow is then transmitted to a centralised traffic control room via a UHF (ultra-high frequency) transmitter (B). The UHF antenna is similar to the aerials used for UHF television reception. The station requires less than 10 W of electrical power and can be easily powered by a solar cell (C) and rechargeable battery system in order to continuously monitor traffic day or night. The self-contained stations are ideal for regional motorways like the Geelong Road. These solar-powered stations are a versatile, efficient and cost-effective way of monitoring traffic flow on our motorways.



Figure 5.20 What is this strange looking device on Melbourne's Geelong Road?

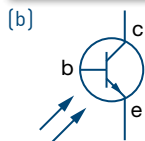
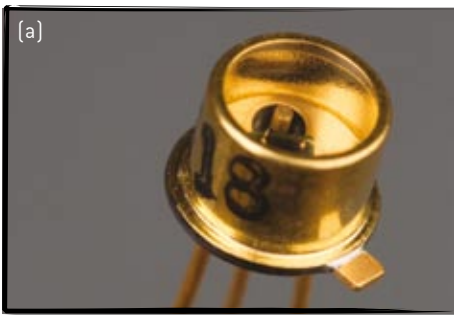


Figure 5.21 (a) A three-terminal phototransistor, although the third (base) terminal is usually left unconnected. (b) A phototransistor circuit symbol (note that some devices have a base terminal whereas others do not).

Phototransistors

Unlike a normal transistor, a phototransistor only needs two terminals—the collector and the emitter. A phototransistor does not require a base terminal, as the base current is generated by light absorbed at the base–emitter depletion region, but in all other respects it essentially operates like a normal transistor. The base photocurrent is amplified by transistor action, and the collector current is proportional to the intensity of light illuminating the device. Because the phototransistor (like a normal transistor) has a high gain (amplification factor), it is generally much more sensitive than a photodiode, although its response is usually slower. Typically, a phototransistor could have a photocurrent gain of between 10 and 100. Phototransistors are suitable for detecting light signals that vary down to timescales approaching $1\ \mu\text{s}$. The time response is inversely related to the gain (i.e. the lower the photocurrent gain, the faster the response time). Two phototransistors are shown in Figure 5.21a. The one on the left has two terminals (collector and emitter), while the one on the right has three terminals, although the third (base) terminal is usually left unconnected in any circuit. Figure 5.21b shows the circuit symbol for a phototransistor. The two simple phototransistor circuits shown in Figure 5.22 can be used to measure light intensity.

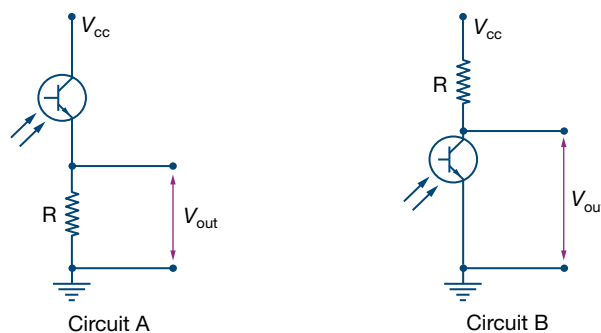
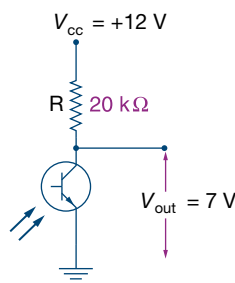


Figure 5.22 Two simple phototransistor detector circuits.



Worked example 5.2C

The phototransistor shown has an active area of $1\ \text{mm} \times 1\ \text{mm}$. It is illuminated with red light of wavelength $620\ \text{nm}$ and intensity $5\ \text{W m}^{-2}$.

- Determine the optical power absorbed by the phototransistor. (Assume that all the light incident on the active area is absorbed.)
- Determine the number of photons per second that are absorbed by the phototransistor.

Solution

- Intensity $(\phi) = \frac{P}{A}$, where P is optical power and A is area.

$$\begin{aligned}\text{Thus } P &= \phi \times A \\ &= 5 \times 1 \times 10^{-6} \\ &= 5\ \mu\text{W}\end{aligned}$$

- Optical power: $P = \frac{\Delta E}{\Delta t}$, where $\frac{\Delta E}{\Delta t}$ is the optical energy absorbed per unit time.

$$P = E_p \frac{\Delta N}{\Delta t}, \text{ where } E_p \text{ is the energy of an individual photon and } \frac{\Delta N}{\Delta t} \text{ is the number of photons absorbed per unit time.}$$

$$\text{Hence } \frac{\Delta N}{\Delta t} = \frac{P}{E_p}$$

$$\begin{aligned}\text{Now } E_p &= \frac{hc}{\lambda} \\ &= \frac{6.626 \times 10^{-34} \times 2.998 \times 10^8}{620 \times 10^{-9}} \\ &= 3.204 \times 10^{-19} \text{ J}\end{aligned}$$

and therefore

$$\begin{aligned}\frac{\Delta N_p}{\Delta t} &= \frac{5 \times 10^{-6}}{3.204 \times 10^{-19}} \\ &= 1.561 \times 10^{13} \text{ photons per second}\end{aligned}$$

Optical sources

Light-emitting diodes

We discovered in Chapter 4 that when a semiconductor pn junction is forward biased, the externally applied voltage reduces the effective barrier potential across the depletion region. This means that electrons and holes can diffuse across the junction very easily, resulting in a significant forward-biased current. When an electron diffuses from the n region into the p region, there is a high probability that it will quickly ‘fill’ one of the many holes in the p region. This is called recombination. (Similarly, holes diffusing in the opposite direction will be quickly ‘filled’ by one of the electrons in the n region close to the depletion layer.) When an electron ‘fills’ a hole, energy is released (because now the atoms near the filled hole have a stable configuration of electrons in their outer shells). The structure and electrical properties of crystalline silicon are such that in a silicon semiconductor diode the energy given off when an electron ‘fills’ a hole is usually converted to heat. If there is a large forward-biased current flowing, you can actually feel the recombination heat generated in the diode.

If, however, a forward-biased pn junction is made of gallium arsenide (GaAs) rather than silicon, then when an electron ‘fills’ a hole there is a high probability that the energy released will be converted to a photon of light (in this particular case near-infrared light). A GaAs diode produces near-infrared radiation with a peak wavelength (λ_{peak}) of around 900 nm and a wavelength range ($\Delta\lambda$) of around 30 nm. This mechanism is shown schematically in Figure 5.23. This type of diode is called a light-emitting diode (LED), and is designed to have its junction close to the surface so that

Physics file

LEDs are becoming very important light sources because they are efficient and reliable. As production costs have decreased, LEDs have replaced many normal light sources. For example, LED torches use considerably less power than a conventional DC light bulb torch. They also require smaller batteries and are much lighter. The LED torch gives off a very bright and evenly diffused bluish white light.

In Australia, traditional ‘globe’ traffic lights are being replaced by LED versions with banks of red, green and amber LEDs. These LEDs are brighter and more reliable than conventional traffic lights. In Singapore, all traffic lights are now the LED type.

Photonics engineers estimate that there would be a huge saving in the world’s energy resources by replacing many of the existing street lights with LED equivalents.

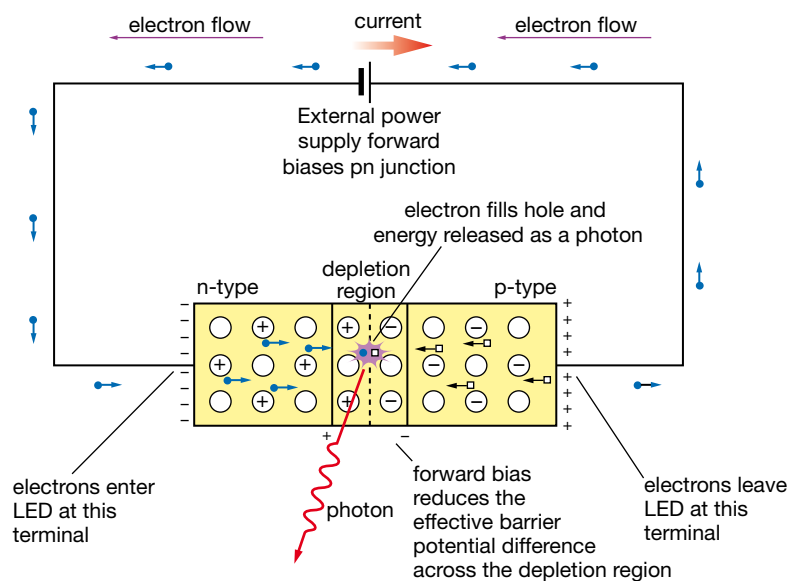


Figure 5.23 A forward-biased LED.

some of the light generated when the diode is forward biased can escape from the device (see Figure 5.24a). The semiconductor surface is usually embedded in a transparent, plastic, hemispherical dome to increase the amount of light escaping (see Figure 5.24b). Different LEDs emit light of different wavelengths, depending upon the particular semiconductor used in their construction. Some different LED semiconductors and the peak wavelengths of the optical radiation that they emit are shown in Table 5.2.

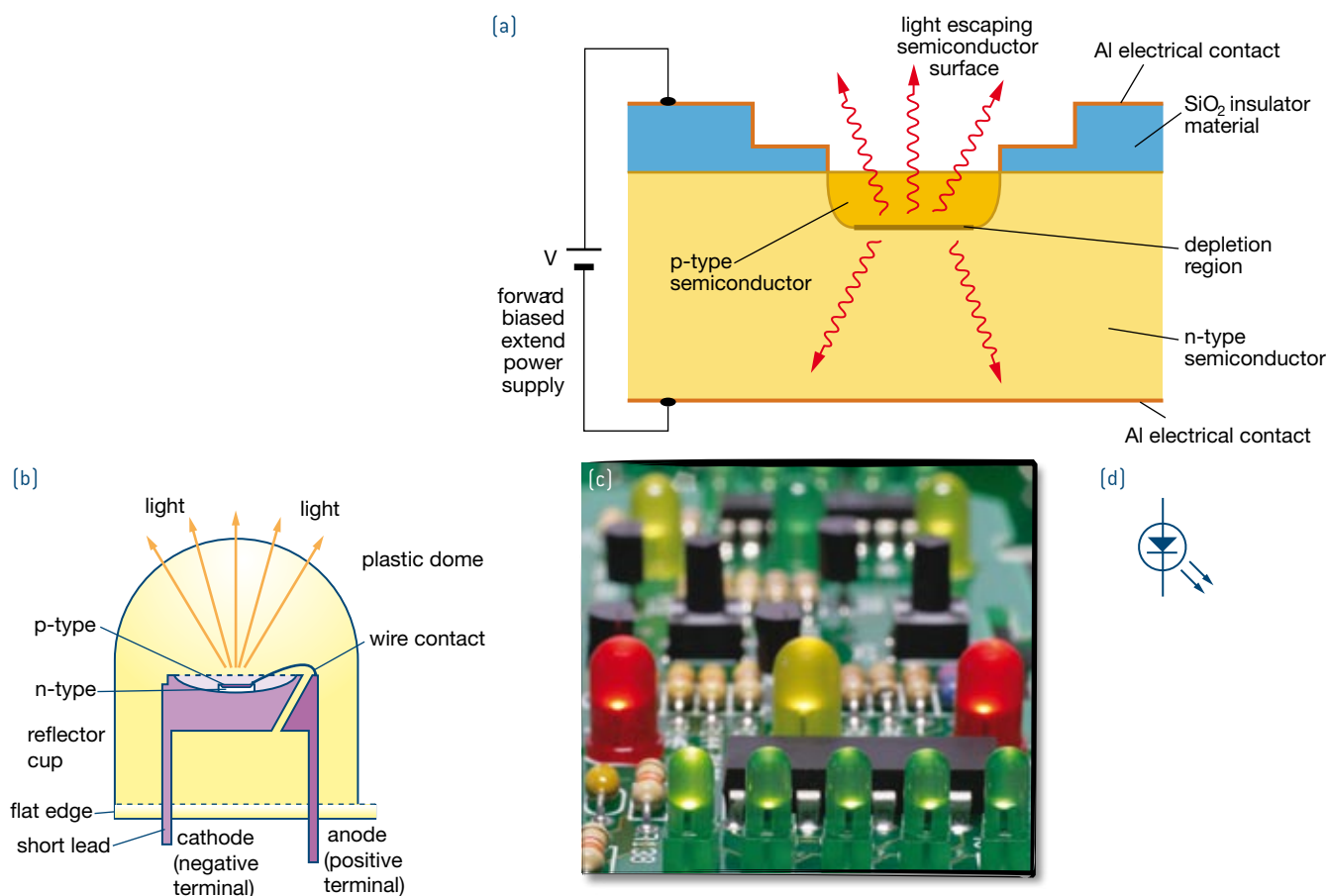


Figure 5.24 (a) Schematic diagram of LED, (b) sketch of LED, (c) various LEDs and (d) LED circuit symbol.

Table 5.2 LED materials, peak wavelengths and colours

Material	λ_{max} (nm)	Colour
InGaAsP*	variable from 1000 to 1550	IR (wavelengths used in modern fibre-optic telecommunication systems)
GaAs	900	IR
GaAsP	665	red
GaPN*	550, 590	green, yellow
GaN	430	blue

*Different peak wavelengths are possible by varying the percentage of the elements making up the LED.

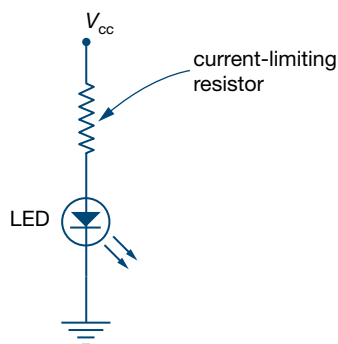


Figure 5.25 LED driver circuit.

The intensity of light emitted by an LED is directly proportional to the forward-biased current flowing through the pn junction. From the I - V characteristics of a diode (see Figure 4.13, page 130), we can see that the forward-biased current (and therefore the LED's light intensity) will increase rapidly once the forward bias reaches the switch-on voltage, V_s , for the particular LED. Many of the common LEDs have a switch-on voltage

of around 1.8–3.5 V. Figure 5.25 shows a typical circuit used to activate an LED. Note that the limiting resistor is used to keep the LED's forward-biased current within the maximum allowed rating. If the current exceeds this limit, the LED will be permanently damaged.

LEDs can also be switched on and off very rapidly (by varying the forward-biased voltage around V_s), often as fast as a fraction of a microsecond.



Figure 5.26 (a) An LED torch and (b) LED traffic light.

Worked example 5.20

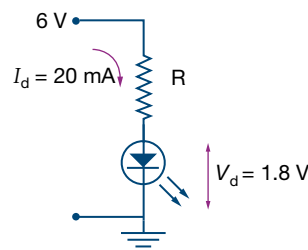
An LED has the following optimum operating characteristics:

$I_d = 20$ mA when $V_d = 1.8$ V

Determine the value of the current-limiting resistor (R) if the current through the diode is to be limited to 20 mA when powered by a 6 V battery.

Solution

$$\begin{aligned} R &= \frac{V_R}{I} \\ &= \frac{6 - 1.8}{20 \times 10^{-3}} \\ &= 210 \, \Omega \end{aligned}$$



Physics in action

Lighting up the world

'Every night 2 billion people in our world are in the dark as soon as the Sun goes down'.

In 1997, Dr Dave Irvine-Halliday visited a Nepalese village where he saw how the lack of lighting was reducing the children's ability to read and learn at night. Over the next few years, he searched for innovative and practical solutions to lighting problems in developing countries. He developed a lighting system based on white-light-emitting diodes (WLEDs), which can light up an entire village on less energy than that used by a single, conventional 100 W incandescent light bulb.

Dr Irvine-Halliday is a photonics engineer, and a professor in the Electrical and Computer Engineering Department of the University of Calgary in Canada. His research interests include the study of photonics devices for telecommunication and sensing, and biophotonics, but Dr Irvine-Halliday now has another interest that he is very passionate about. He is the President of the Light Up The World (LUTW) Foundation—a non-profit organisation that is helping many poor and isolated villages in developing countries. He established LUTW to develop a low-cost, safe, healthy, reliable and very efficient WLED lighting system and to facilitate its introduction throughout developing countries (see Figure 5.27).



Figure 5.27 A Sri Lankan mother and child holding an LUTW WLED lamp. This lamp, which has been successfully used in many developing countries, has nine 0.1 W WLEDs manufactured by Nichia of Japan, the inventors of the original WLED.

Typically, the WLED lamps are powered by rechargeable batteries that use renewable energy sources (solar cells, pedal generators, miniature (pico) wind or hydro generators etc.). The US company Lumileds, which makes the 1 W WLEDs, supplies them in bulk to LUTW for only a few dollars each. Currently, the WLED home lighting systems, including lamps, battery and solar cell, cost around \$US40. This is a one-time cost that gives the household effectively free light until the battery requires replacing, which is expected to be at least 5 years. Over the next few years, one of LUTW's top priorities is to lower the one-time cost to \$US20; that is '20-20 Vision', or \$US20 for 20 years of home lighting.

Dr Irvine-Halliday's most recent solid state lighting systems are designed around a 1 W state-of-the-art WLED that generates light 2.5 times more efficiently than an incandescent light globe. The WLED lamp has a further advantage over an incandescent light globe since it directs all its light in the direction where it is most useful. Dr Irvine-Halliday has found that 'a single 1 W WLED lamp produces ample lighting for four children to read and study by at a moderate-sized desk'. His latest model uses the 1 W Luxeon WLED manufactured by Lumileds of the USA and produces enough light for an entire household's living area.

Before LUTW, isolated villages in developing countries relied on candles and kerosene wick lamps for their lighting, which together with wood cooking and heating resulted in high indoor smoke pollution and serious respiratory problems. The WLED lighting system, on the other hand, provides a safe, healthy, affordable and very reliable light source.

So far, LUTW and its affiliates have permanently lit the homes of many thousands of villagers in Nepal, India, Sri Lanka, Dominican Republic, Haiti, Guatemala, Bolivia and Irian Jaya, and its Luxeon lamps are being tested in villages in Tibet, Nicaragua, Brazil, Chile, Peru, Angola and Uganda. To find out more about the LUTW Foundation, visit their website.

Laser diodes

A laser diode (LD) works on exactly the same principle as the LED (i.e. the same mechanism where forward biasing a pn junction results in recombination of holes and electrons which, for certain semiconductor materials, leads to the emission of light rather than heat). The main differences are that in a laser diode, the semiconductor material has a special current-confining geometry and very high dopant concentrations.

This means that the light emitted from a laser diode has a much narrower range of wavelengths (typically 1–5 nm) than an LED (typically 30–80 nm) and it is emitted in a narrower beam than most LEDs. The LD light can also be switched on and off in a fraction of a nanosecond, which is much faster than the switching rates possible with an LED.

Laser diodes, like LEDs, can convert a change in the diode current *directly* into a change in light intensity. The change in light intensity is directly proportional to the change in current and this can be used as a method of imprinting information onto the light beam in laser telecommunications applications. This process is known as intensity (or amplitude) modulation, as the information signal modulates (or directly changes) the intensity of the laser diode's light output.

Laser diodes are much more temperature sensitive than LEDs and often require more complicated electronic drive circuits. They sometimes use optical feedback from an inbuilt photodiode to maintain a constant light intensity output (see Figure 5.28 for some examples of LDs). Figure 5.29 shows the circuit symbol for a laser diode.

Physics file

Laser stands for: **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation.

Other properties of lasers are covered in more detail in Chapter 14.

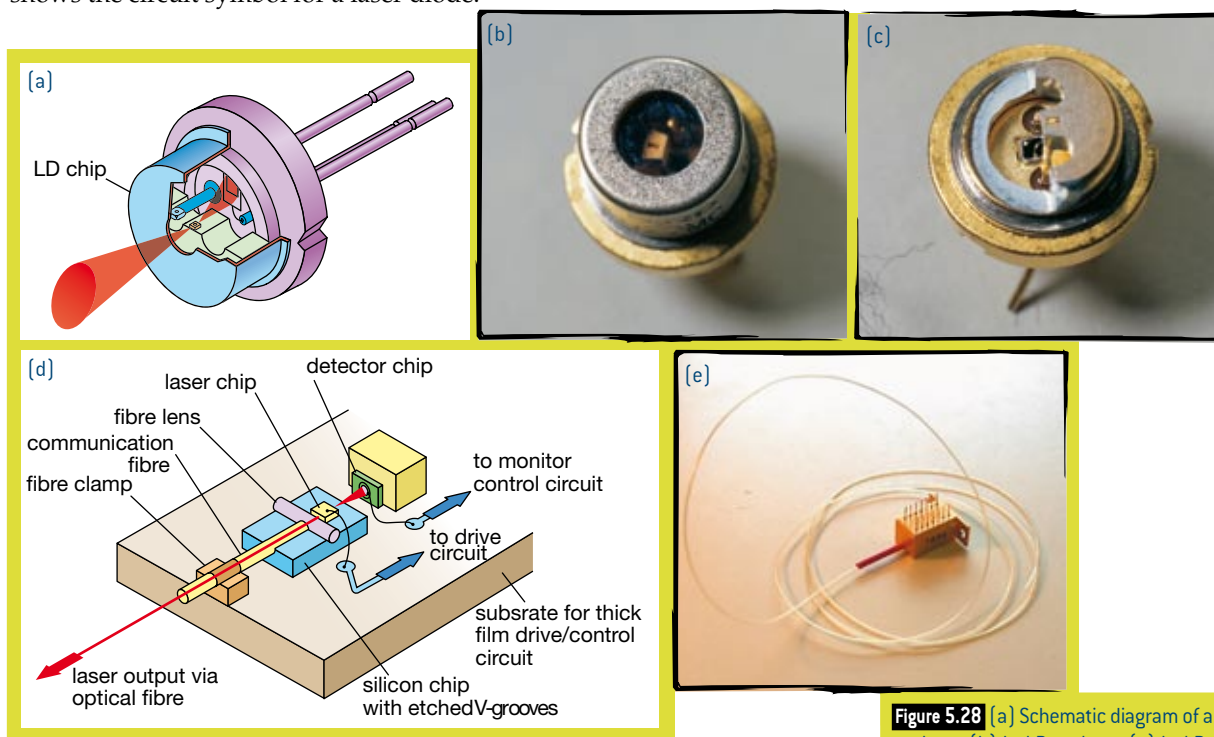


Figure 5.28 (a) Schematic diagram of an LD package. (b) An LD package. (c) An LD package with the cover removed, so that the active area of the device can be clearly seen. (d) Schematic diagram of LD connected directly to an optical fibre [this device is called a pigtailed LD]. (e) An LD pigtail.

We are all aware of the role of LDs in laser pointers, barcode readers, and CD and DVD players. But LDs also play a very important (but essentially unseen) role in fibre-optic telecommunication systems, medical lasers and 'smart structure' systems for civil and aerospace applications etc. The development of new blue laser diodes is expected to open up other areas of application for LDs (e.g. biotechnology, fabrication of nano-scaled devices).

Worked example 5.2E

The simplified circuit shown on page 168 is used to modulate the output of the laser diode with a small time-varying input voltage signal (Δv_{in}). The laser diode should be operated at its optimal DC operating conditions (i.e. $I_d = 10$ mA when $V_d = 2.2$ V). Determine the values of R_1 , R_2 and R_3 , which bias the transistor in the middle of its operating range and allow the laser diode to operate at its optimal conditions.

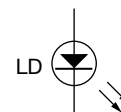
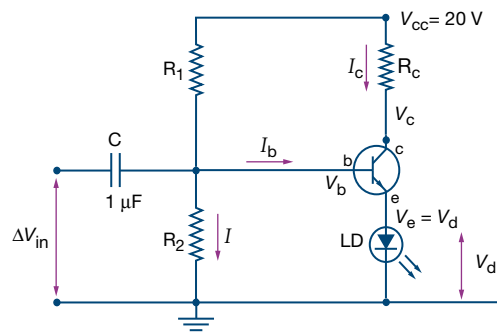


Figure 5.29 Circuit symbol for a laser diode.



Assume that the transistor has a current gain of 100.

Hint: It may be useful to revise transistor amplification and biasing in Chapter 4.

Solution

Since $I_d = I_e \approx I_c$, we require $I_c = 10 \text{ mA}$.

To operate the transistor in the middle of its range, we require:

$$\begin{aligned} V_c &\approx \frac{1}{2} V_{cc} \\ &= \frac{1}{2} \times 20 \\ &= 10 \text{ V} \end{aligned}$$

$$\text{and } V_e = \frac{1}{10} V_{cc}$$

Fortunately $V_e = V_d = 2.2 \text{ V}$, which is approximately equal to $\frac{1}{10} V_{cc}$.

$$\begin{aligned} \text{Hence } R_c &= \frac{V_{cc} - V_c}{I_c} \\ &= \frac{20 - 10}{10 \times 10^{-3}} \\ &= \frac{10}{10 \times 10^{-3}} \\ &= 1 \text{ k}\Omega \end{aligned}$$

$$\text{and } I_b = \frac{1}{100} I_c = 0.1 \text{ mA [since current gain is 100].}$$

For correct operation of the transistor and R_1/R_2 voltage divider, we want the current through R_2 (I) to be at least a factor of 10 greater than I_b ; hence:

$$\begin{aligned} I &= 10 I_b \\ &= 10 \times 0.1 \times 10^{-3} \\ &= 1 \text{ mA} \end{aligned}$$

But:

$$\begin{aligned} V_b &= V_{be} + V_d \\ &= 0.7 + 2.2 \\ &= 2.9 \text{ V} \end{aligned}$$

Hence:

$$\begin{aligned} R_2 &= \frac{V_b}{I} \\ &= \frac{2.9}{1 \times 10^{-3}} \\ &= 2.9 \text{ k}\Omega \end{aligned}$$

We can now calculate the value of R_1 from the R_1/R_2 voltage divider:

$$\begin{aligned} V_b &= V_{cc} \left(\frac{R_2}{R_1 + R_2} \right) \\ 2.9 &= 20 \left(\frac{2.9 \times 10^3}{R_1 + 2.9 \times 10^3} \right) \\ 2.9 R_1 + [2.9 \times 2.9 \times 10^3] &= 20 \times 2.9 \times 10^3 \\ 2.9 R_1 &= [20 - 2.9] \times 2.9 \times 10^3 \\ 2.9 R_1 &= 17.1 \times 2.9 \times 10^3 \\ R_1 &= 17.1 \times 10^3 \\ R_1 &= 17.1 \text{ k}\Omega \end{aligned}$$



5.2 summary

Optical transducers

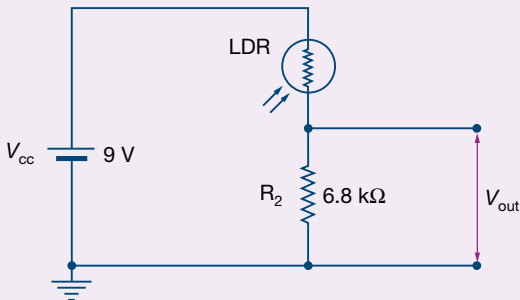
- Light-dependent resistors (LDRs) are semiconductors whose resistance depends upon illumination. Their response is non-linear and slow.
- Photodiodes are pn-junction semiconductor devices which, in photoconductive mode (i.e. when reverse biased), generate a photocurrent that depends upon illumination. The response is linear and fast.
- Phototransistors are BJT devices that, when biased correctly, can generate a collector current that depends upon illumination. They have a high optical gain. The response is linear, and the response time moderate.
- The light-emitting diode (LED) is a pn-junction device and the intensity of its emitted light is directly proportional to its forward-biased current. Its light output can be modulated rapidly, and it has a narrow range of emission wavelengths.
- The laser diode (LD) is a pn-junction device and the intensity of its emitted light is directly proportional to its forward-biased current. The device has special current-confining geometry and very high dopant levels, and emits *laser light* (when the current is above a given threshold). Its output can be modulated rapidly, and it has a very narrow range of emission wavelengths.
- The energy of a discrete photon is given by $E_p = hf = \frac{hc}{\lambda}$, where f is frequency, λ is the wavelength, c is the speed of light in a vacuum and h is Planck's constant.

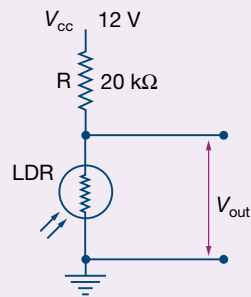


5.2 questions

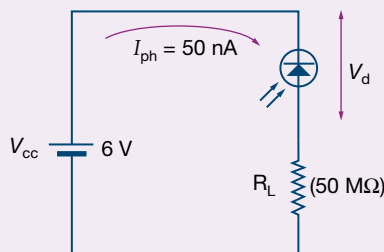
Optical transducers

- The light emission from an LED can be varied by:
 - varying the current passing through the diode
 - varying the illumination of the diode
 - varying the reverse-biased voltage across the diode
 - all of the above.
- Which one of the following statements is true about the difference between LDs and LEDs?
 - LDs always emit more light than LEDs at the same drive current.
 - LDs can be directly modulated whereas LEDs cannot.
 - LDs emit light over a narrower range of wavelengths than LEDs.
 - The output of LDs spreads out over a broader range of angles than that of LEDs.
- Write down one advantage and one disadvantage of using a phototransistor instead of a photodiode in an optical detection circuit.
- When the LDR shown in the diagram is in darkness, it has a resistance (R_{LDR}) of 2 M Ω .


- Calculate the total resistance of the series circuit when the LDR is in darkness.
- Calculate the voltage drop across R_2 (V_{out}) when the LDR is in darkness.
- The LDR is now illuminated with a light source and its resistance decreases. Determine the resistance of the LDR if V_{out} now is 6 V.
- An LDR is used to determine whether the light intensity is adequate for a person working at a desk in an office. The LDR has the characteristics shown in Figure 5.13, and the voltage divider circuit shown is used.


- Determine whether the light level at the desk is above the minimum acceptable level (approximately 100 mW m⁻²) if the output voltage of the circuit is 4 V.
- What is the intensity of the light radiation at the desk?
- Consider the following circuit. If the supply voltage (V_{cc}) is 6 V and the photocurrent (I_{ph}) is 50 nA, determine:
 - the voltage across the photodiode

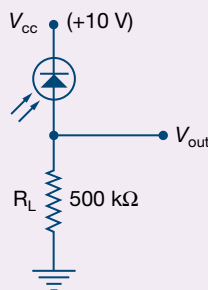
- b** whether the photodiode is operating in photoconduction or photovoltaic mode.



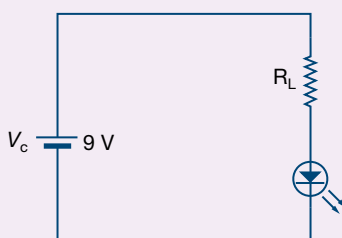
- 7** The photodiode shown in the following circuit has the same I - V characteristics as those shown in Figure 5.16. Assume that the photodiode is to be operated in photoconductive mode.

- a** Determine the maximum current that can flow through R_L .

- b** Determine the maximum light intensity that can be reliably measured by the circuit.



- 8** The LED in the following circuit has a switch-on voltage (V_s) of 2.0 V and an operating current of 20 mA. Determine the value of R_L for the LED to be operating correctly.

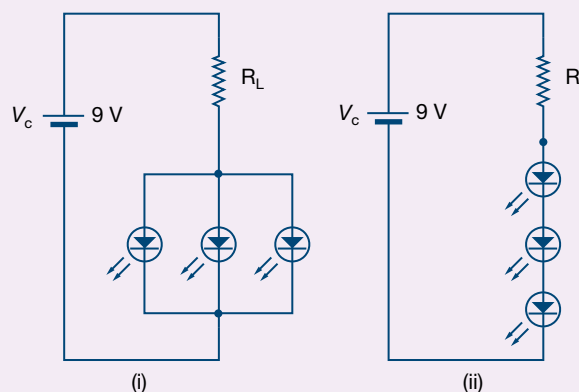


- 9** All LEDs in the following circuits (i) and (ii) are identical. Each LED has a switch-on voltage (V_s) of 2 V and draws a current of 20 mA for optimal light production. (If the current is much smaller, the LED light output is too dim. If the current is much larger, the LED overheats and burns out.)

- a** For each circuit determine the R_L that gives optimum operation for all LEDs.

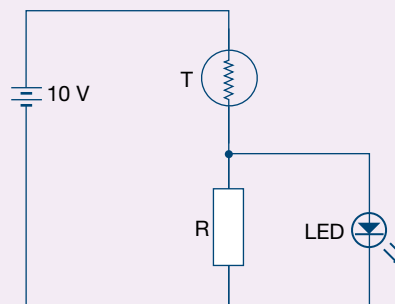
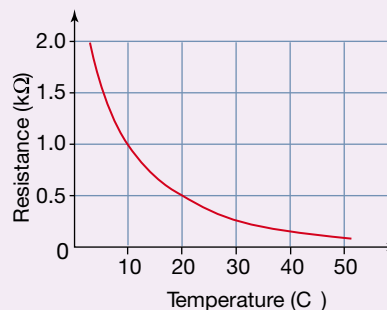
- b** Which combination of LEDs ((i) or (ii)) emits the most light with these particular values of R_L ?

- c** Assuming circuits (i) and (ii) have exactly the same ideal battery power supply, determine which circuit will emit light (i.e. LEDs operating at optimum level) for the longest time. Explain your answer.



- 10** The following graph shows the resistance-temperature characteristics of a thermistor. A circuit uses this thermistor as part of a temperature sensor which can activate an LED whenever the temperature rises above a certain level.

- a** What is the resistance of the thermistor at 20°C?
- b** The potential difference across the LED at 20°C is 2.5 V and the current through it is 11 mA. What is the value of R ?
- c** The LED is activated by a minimum potential difference of 2.0 V across it which gives a current through it of 4.8 mA. What is the minimum temperature that will activate the LED?



5.3 Audio transmission via a light beam

Designing a simple telecommunications system

In this section, we shall design a simple, inexpensive laser-based ‘line-of-sight’ telecommunications system capable of transmitting an audio signal with remarkable fidelity. In a sense, we will rediscover Alexander Graham Bell’s photophone but with a few modern enhancements. Our system will use a laser diode as an extra-bright light source and a semiconductor optical transducer as an extra-sensitive detector.

We can use an inexpensive laser pointer as a suitable laser light source. The laser diode unit found in a typical laser pointer (see Figure 5.30) is a low-powered ($\sim 1\text{ mW}$) device that emits a red ($\sim 670\text{ nm}$) laser beam.

However, take care, as even low-power laser pointers can cause serious eye injury if used incorrectly. Read the following Physics in action on laser safety before proceeding.



Figure 5.30 Laser diode pointer. These are restricted items and should only be used under supervision.

Physics in action

Laser safety

Look at the diagram of the eye in Figure 5.31. The cornea is the clear film that covers the surface of the eye. The iris controls the amount of light entering the eye. The lens is one of the elements that focus the incoming light onto the retina, where light-sensitive cells detect the light. Ultraviolet light tends to be absorbed by the cornea. Although corneal cells can regenerate to an extent, UV radiation in sunlight can permanently damage the cornea if the eye is exposed to excessive bright sunlight over many years. This is one reason for wearing UV-protecting sunglasses during summer. Visible and near-infrared optical radiation is focused onto, and absorbed by, the retina. Exposure to high-intensity light from a source of this optical radiation will cause instantaneous and permanent damage to the retina because the light source will be focused onto a small number of cells and because retinal cells do not regenerate.

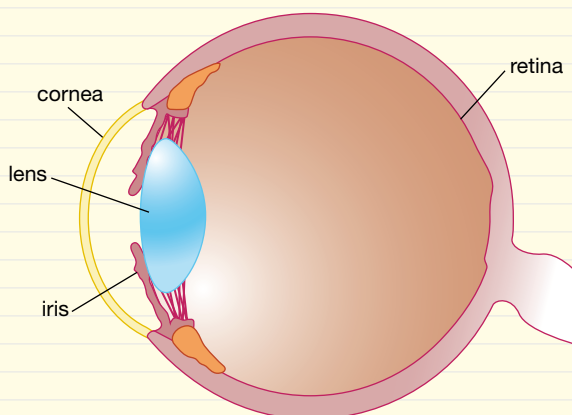


Figure 5.31 Schematic diagram of the eye.

With a normal light source (like an incandescent light bulb), the optical energy spreads evenly in all directions. If you double the distance between your eye and the bulb, then the optical energy entering your eye falls by a factor of four—hence distance from the light source offers a level of protection. With a laser, the optical energy is contained in a very small beam that does not spread out much. Even at relatively large distances, the laser beam spot can still be very small and hence the optical energy entering the eye can still be very high. In addition, the almost parallel laser light irradiating the eye forms a very sharp point focus on the retina. This means that most of the energy emitted by the laser will be concentrated onto only a few cells on the retina, causing maximum damage to those cells. Looking directly into the beam of a low-power 1 mW visible laser can cause considerably more retinal damage than looking directly at the Sun!

Exposure risk can be minimised when working with a laser (in this case a laser pointer) by always observing the following safety procedures.

- Terminate the beam at the end of its useful path (i.e. with a photodetector and/or a non-reflective beam stop behind the detector). A large piece of dark-coloured cardboard mounted vertically in a slotted wooden base makes a good beam stop.
- Avoid specular reflection from polished or shiny surfaces, including watches and jewellery, as these will reflect the laser beam around the room in an uncontrolled manner.
- Align the laser beam with the lowest intensity practicable. Use partially crossed linear polarisers to temporarily reduce the beam intensity. The two polarised plastic lenses of cheap polaroid sunglasses can be used.
- Never look directly into the laser beam. Construct your laser beam path as low as practical to the table or bench top, and then always try to keep your head well above the level of the laser beam. When aligning, move your eye slowly when near

(a)



(b)



(c)

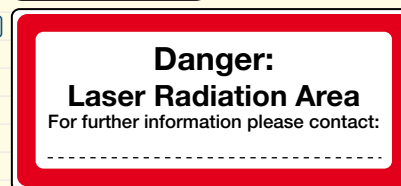


Figure 5.32 (a) Specific laser diode radiation warning sign. (b) General laser radiation warning sign. (c) Contact information sign.

the beam path. The bright halo associated with the beam gives an early warning of impending disaster, as your eye approaches a direct line of vision with the beam.

- Avoid darkened rooms. Do not set up your laser beam in a darkened room, as your pupils will enlarge and potentially let more laser beam energy onto the retina.

- Always use laser warning signs to alert others to the potential laser radiation hazard. You should print enlarged copies of laser radiation warning signs like those shown in Figure 5.32 and display them prominently, close to any laser experiment.



PRACTICAL ACTIVITY 22

Wave division multiplexing

As for the original photophone, we need to produce an intensity-modulated signal (i.e. the intensity of the laser beam should vary in accordance with the fluctuations in the amplitude of the audio signal). We can achieve this by gluing a small piece of opaque plastic or cardboard sheet onto a speaker, so that it can partially interrupt the light beam as the 'transmitter' speaker vibrates (in response to the original audio signal).

The intensity-modulated laser light beam can then travel a large distance to the receiver (carrying the audio information coded as light intensity variations in the beam). The optical detector at the receiver can be a simple voltage divider circuit using an LDR and a suitable fixed resistor. The resistance of the LDR will depend on the changing light intensity illuminating it, and the output voltage of the LDR voltage divider circuit will provide an electrical signal that will be roughly proportional to the original audio signal. The small output voltage from the LDR voltage divider circuit will need to be amplified before being sent to the 'receiver' speaker. A schematic diagram of our simple intensity-modulated laser telecommunication system is shown in Figure 5.33.

There are two limitations that affect the fidelity of our simple laser telecommunication system. First, the small mass of the opaque strip (that we use to change the intensity of the transmitted laser light beam) affects the vibration of the 'transmitter' speaker and greatly attenuates any high-frequency vibrations. Second, the time response of the LDR is slow and hence the optical detection circuit also attenuates any high-frequency components of the transmitted signal. Although the sound from the receiver is a reasonable copy of the original audio signal, there will be some loss of high frequencies and the sound signal will be a little distorted.

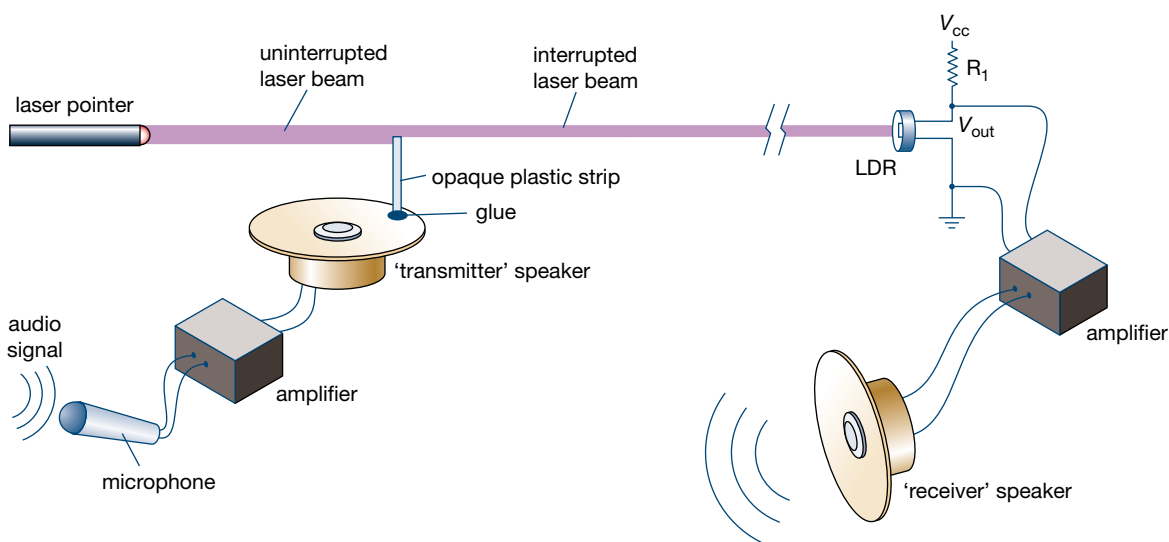


Figure 5.33 Simple intensity-modulated laser telecommunication system.

Physics file

Warning: Battery-powered laser pointers are restricted items. Use only under teacher supervision.

Designing a high-fidelity photonics telecommunications system

To overcome the first limitation of our original telecommunications prototype, we need to use a better modulation system. Ideally, we would like to convert the electrical signal from the audio source (microphone, cassette tape player, radio, CD player etc.) *directly* into a change in the light intensity output of the laser diode. In other words, we would like to intensity-modulate our laser beam directly via the electronic signal from our audio source.

The laser pointer comprises a laser diode package, focusing lens assembly and drive circuitry all interconnected as one LD unit. If the barrel casing of the laser pointer could be removed (usually a difficult task) the LD unit would look similar to the one shown in Figure 5.34. The photo clearly shows

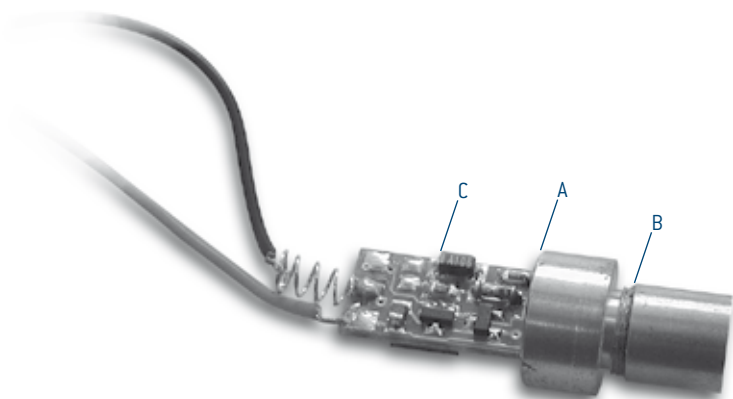


Figure 5.34 Laser diode unit inside a laser pointer: A, laser diode package; B, focusing lens assembly; C, drive circuitry.



Figure 5.35 Laser pointer with barrel casing partially cut away.

Physics file

If the laser diode module is not well heat-sinked, the operating temperature of the LD's pn junction will be higher than normal. Under these conditions, the LD's performance will degrade and its lifetime decrease. This effect can be minimised by operating the LD below its normal output power.

the LD device, the focusing lens assembly and the drive circuit. The size of the laser spot can usually be fine-tuned by adjusting the focusing lens assembly. The LD unit is usually powered by a 4.5 V source, which in the case of the laser pointer is three small 1.5 V button batteries connected in series. To modulate the output of the laser pointer directly, we first need to remove the batteries and replace them with an external power supply.

The aluminium barrel casing of the laser pointer can be partially cut away with a pair of small tin snips as shown in Figure 5.37. Take care to smooth any sharp edges on the cut casing. A green plastic-coated copper wire can then be connected to the central spring of the drive circuit, which is the negative connection. The wire can be connected either with solder (as shown in the photo) or with alligator clips (that have an insulated cover). A red plastic-coated copper wire can also be connected to the barrel case, which is the positive connection of the drive circuit.

The laser pointer's push button switch can be permanently connected in the 'on' position by wrapping a rubber band or some adhesive tape around the barrel of the pointer. The pointer can then be turned on or off via the external power supply without disturbing the alignment of the laser beam.

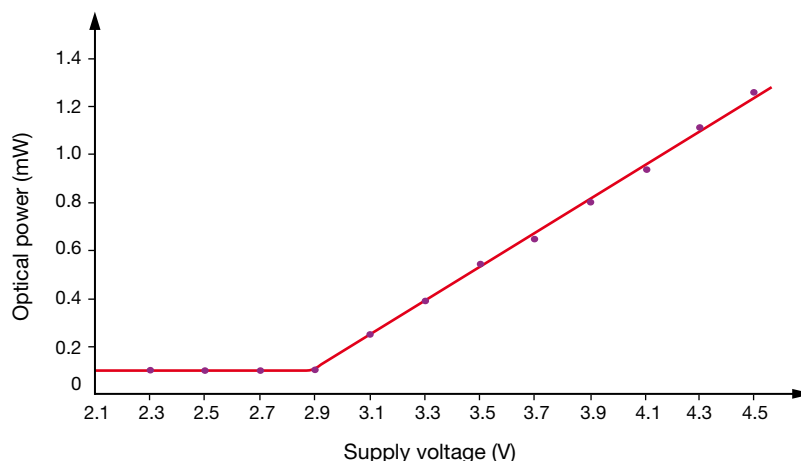


Figure 5.36 Optical power versus supply voltage for the laser pointer.

Figure 5.36 shows a plot of optical power output of the laser pointer measured as a function of voltage supplied to the drive circuitry. As can be seen, above the threshold of 2.9 V, the light output of the laser diode (which is directly proportional to the current through the diode) is essentially linearly proportional to the supply voltage. So if we can superimpose a voltage representing the audio signal onto the power supply voltage, we will be able to modulate the light intensity *directly* from the audio signal. We need

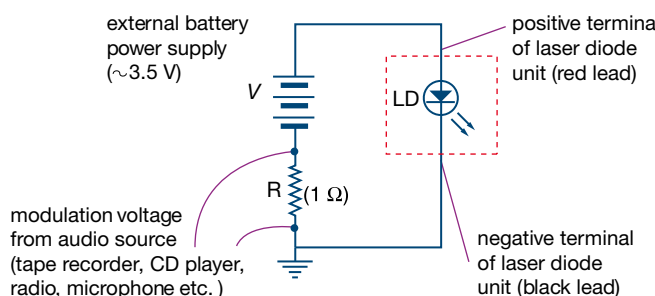


Figure 5.37 Schematic diagram of a laser transmitter with direct intensity modulation.

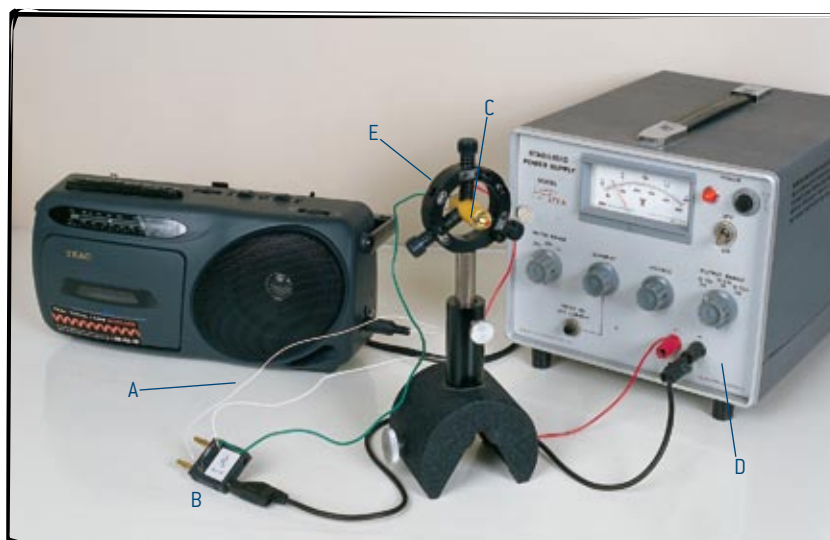


Figure 5.38 Prototype of the laser transmitter with direct intensity modulation: A, line out of a cassette recorder; B, $1\ \Omega$ resistor; C, laser pointer; D, external variable power supply; E, optical support frame.

Physics file

The outer case of many (but not all) laser diode modules is at a positive potential.

If you are unsure about the polarity of any particular LD, check the manufacturer's specifications.

to be aware that our supply voltage for the drive circuit should not go above its normal operating voltage, in this case 4.5 V, as this may produce an optical power that exceeds the laser diode's specifications and might result in permanent damage to the unit.

Figure 5.37 is a schematic diagram of our laser transmitter with direct intensity modulation. A small $1\ \Omega$ resistor has been inserted in series with the power supply and the terminals of the LD unit. The addition of the $1\ \Omega$ resistor has very little effect on the operating characteristics of the LD unit, but does allow a small voltage signal (from an audio source) to be superimposed on the supply voltage, thereby allowing the laser output to be modulated. The electrical signal from the *line out* or *headphone* output from a tape recorder, CD player, radio or other audio source can be connected across the $1\ \Omega$ resistor as shown in Figure 5.37. Make sure that your circuit has only one earth (ground) point, which is usually the negative lead of the *line out*. This means that your external power supply (or external battery if you wish) should *not* be separately earthed.

It is often a good idea to operate the laser diode below its normal output power, by using a slightly lower supply voltage. This reduces the temperature of the laser diode's pn junction and greatly increases its lifetime.

Figure 5.38 shows an example of the prototype laser transmitter with direct-intensity modulation. In this particular case, the *line out* of a cassette recorder was connected across the $1\ \Omega$ resistor. The resistor was connected in series with the terminals of the laser pointer and the outputs of an external variable power supply. Neither of the power supply terminals (positive or negative) was grounded (connected directly to earth). The output of the power supply was adjusted to give approximately 3.2–4 V DC. The varying voltage (audio signal) across the $1\ \Omega$ resistor (which is superimposed on the DC voltage) was approximately 100–200 mV (peak to peak). The laser pointer was mounted in an optical support frame with plastic insulating holding screws. The insulating screws are needed because the barrel of the laser pointer is at a positive potential and we need to prevent the barrel

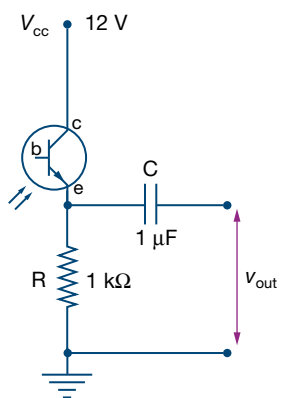


Figure 5.39 Schematic diagram of the phototransistor receiver.

accidentally shorting to earth. A retort stand with insulated clamps makes a suitable substitute for the optical support frame. The laser pointer must be adjusted so that the laser beam is roughly horizontal to the bench surface.

In the prototype, a phototransistor rather than an LDR is used as the optical detector. The phototransistor has a much better high-frequency response than the LDR and gives a much larger output voltage because of its internal optical amplification. Figure 5.39 is a schematic diagram of the phototransistor detector circuit. When illuminated by the laser beam, the phototransistor circuit (in this particular prototype) gave an output voltage for the audio modulation of approximately 1 V (peak to peak). The DC voltage component of the signal is blocked by the 1 μ F coupling capacitor. The audio component of the signal (modulation) must then be amplified before being connected to the 'receiver' speaker.

Figure 5.40 shows an old prototype laser receiver that could now be readily built in the school lab. The phototransistor, resistor and capacitor are all connected on an electronic breadboard. The breadboard is held in a retort stand that is aligned so that the laser beam illuminates the phototransistor. The variable power supply provides the 12 V to power the phototransistor detector circuit. The output of the detector circuit goes to an inexpensive

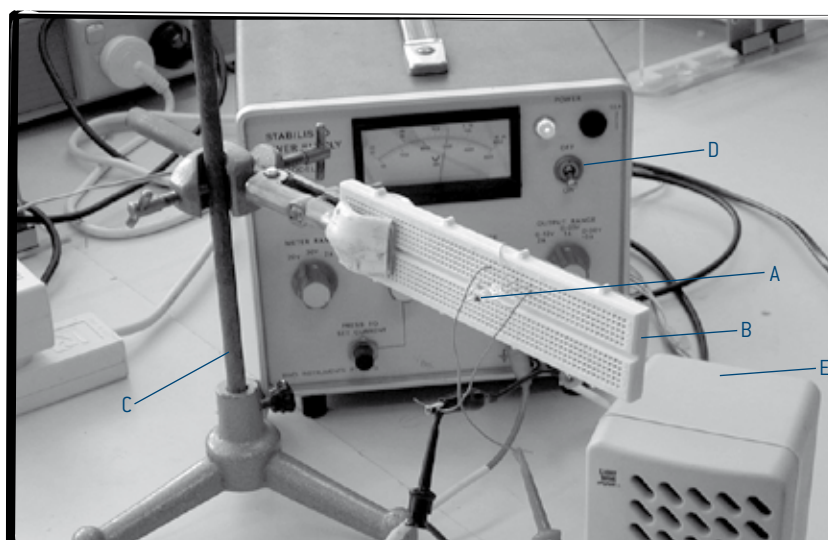


Figure 5.40 Prototype laser receiver: A, phototransistor; B, electronic breadboard; C, retort stand; D, variable power supply; E, amplified computer speaker.



Figure 5.41 Phototransistor with black collimation tube.

amplified computer speaker. An infrared phototransistor was used in this prototype, but it had sufficient sensitivity to detect red light. The longer lead is the emitter and the shorter lead is the collector (it has no base terminal). Phototransistors are readily available from many electronics stores. A black cardboard or plastic collimator tube (as shown in Figure 5.41) can be made to fit over the head of the phototransistor. This will reduce the amount of stray light incident on the detector.

The audio reproduction from this prototype laser telecommunication system is remarkably good, given the low cost of the components, and the ease with which the system can be constructed. Essentially it shows the main features (light transmitters, modulation, optical receivers) found in modern fibre-optic telecommunication systems. Photonics telecommunication systems are growing very rapidly, as they are the enabling technology of the Internet.

Although the fibre-optics telecommunications is a very important application of photonics, the science of photonics covers a broader range of topics. These topics are discussed in Chapter 14.



5.3 summary

Audio transmission via a light beam

- Intensity (or amplitude) modulation is when the intensity of a light beam varies in accordance with the fluctuations in the amplitude of the information signal.




5.3 questions

Audio transmission via a light beam

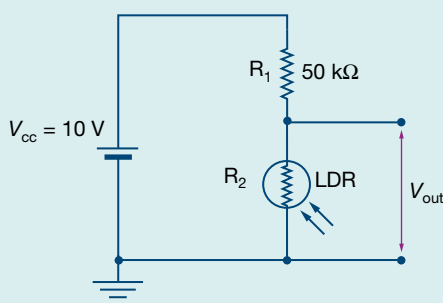
The following questions refer to the laser telecommunication system developed in section 5.3.

- Which modulation method was used for the laser telecommunication system that transmitted the audio signal?
 - intensity modulation
 - frequency modulation
 - wavelength modulation
 - sound modulation
- With the laser telecommunications system used to transmit an audio signal, why is an LD used in preference to an LED?
 - LEDs emit a wider range of wavelengths than LDs.
 - Unlike an LED, an LD's beam spot does not spread out much, even over long distances.
 - Unlike an LED, an LD's light output can be easily modulated.
 - All of the above.
- With the laser telecommunication system used to transmit an audio signal, why is it preferable to use a phototransistor rather than an LDR?
 - The phototransistor generates a larger output voltage for the modulation.
 - The LDR has a poorer high-frequency response.
 - The phototransistor uses 'transistor action' to amplify the signal illuminating its base.
 - All of the above.
- What is the most significant limitation associated with using an opaque object to interrupt the laser beam to create modulation of the audio signal?
 - Difficulty with permanently gluing object to speaker membrane
 - Loss of high-frequency response of speaker when object is attached
 - Movement of speaker membrane insufficient to interrupt all the laser beam
 - Opaque object heats up when irradiated by the laser beam
- With the laser pointer used in the laser telecommunication system, the positive terminal of the external battery or power supply:
 - should be connected only to the central spring of the drive circuit

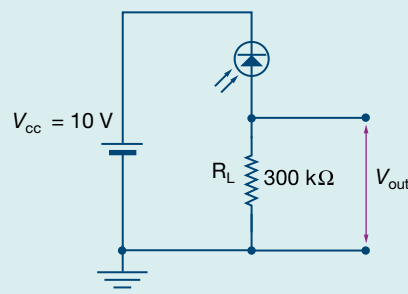
- 
- B** can be connected to either the central spring or the outer barrel casing
- C** should be connected only to the outer barrel casing
- D** should be connected to neither the central spring nor the outer barrel casing.
- 6** The output of the laser pointer used in the laser telecommunication system varies as a function of the power supply voltage. Which one of the following is true?
- A** The output varies linearly between 0 and 2.9 V.
- B** The output varies linearly between 0 and 4.5 V.
- C** The output varies linearly between 4.5 and 12 V.
- D** The output varies linearly between 2.9 and 4.5 V.
- 7** If the power supply to the laser diode pointer was increased to 12 V, which one of the following would occur first?
- A** The laser diode would have too much current flowing through it and the diode would be permanently damaged.
- B** The laser diode would draw too much current from the power supply and the supply would be permanently damaged.
- C** The laser diode would emit too much light and the receiver speaker would be permanently damaged.
- D** The laser diode would emit too much light and the optical detector would overheat.
- 8** The laser diode pointer should be mounted into an insulated supporting frame so that:
- A** the laser diode's outer case is not accidentally shorted to earth
- B** the experimenter only handles the frame and not the laser diode, thus avoiding electrocution
- C** the supporting frame does not overheat
- D** all of the above.
- 9** The power supply for a laser diode should never exceed its normal operating value by more than:
- A** 500%
- B** 0%
- C** 50%
- D** 100%
- 10** The 0.1 μF capacitor in the optical phototransistor detector circuit:
- A** reduces the 'transistor action' of the detector
- B** increases the 'transistor action' of the detector
- C** filters out any DC contribution of the laser beam
- D** none of the above.

chapter review

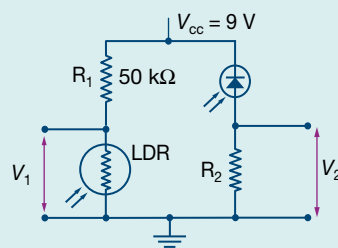
- Photodiodes, as used in fibre optic telecommunication systems, are normally:
 - forward biased to generate a voltage that is directly proportional to the light intensity
 - reverse biased to generate a current that is directly proportional to the light intensity
 - forward biased to generate a current that is directly proportional to the light intensity
 - unbiased to generate a voltage that is directly proportional to the light intensity.
- Which one of the following is generally true for the response times of photodetectors?
 - LDRs are faster than phototransistors which are faster than photodiodes.
 - Phototransistors are faster than LDRs which are faster than photodiodes.
 - Photodiodes are faster than LDRs which are faster than phototransistors.
 - Photodiodes are faster than phototransistors which are faster than LDRs.
- Which one of the following is suitable for a high-speed fibre-optics telecommunication system?
 - A receiver which uses an LDR.
 - A receiver which uses a reverse-biased photodiode.
 - A transmitter which uses a reverse-biased LD.
 - A transmitter which uses a forward-biased phototransistor.
- An LED draws 100 mA when 2 V is applied across its terminals. Under these conditions, the LED produces an optical power of 2 mW.
 - How much electrical power is dissipated in the LED?
 - What is the LED's conversion efficiency from electrical to optical power?
- A photodiode has a circular active area with a radius of 2 mm. Determine the light intensity (W m^{-2}) at the detector if the optical power detected is 0.1 mW.
- The output of the circuit shown is $V_{\text{out}} = 3 \text{ V}$. If the LDR has the same characteristics as those shown in Figure 5.13, determine the intensity of light illuminating the detector.



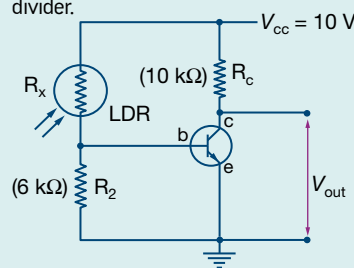
- If $V_{\text{out}} = 6 \text{ V}$, determine the light intensity illuminating the photodiode in the following circuit, assuming it has the same I - V characteristics as those of Figure 5.16.



- The circuit shown has an LDR detector (V_1) and a photodiode detector (V_2). The LDR and the photodiode are both illuminated with light of the same intensity (ϕ). The LDR and the photodiode have the same characteristics as those shown in Figure 5.13 and Figure 5.16, respectively.



- If the intensity of the light increases ($\phi_{\text{final}} > \phi_{\text{initial}}$), determine what happens to V_1 and V_2 (increase, decrease, no change). Give reasons for your answer.
 - If the light intensity (ϕ) is 3 W m^{-2} , calculate V_1 .
 - The dual detection system is to be calibrated so that both outputs are the same for a light intensity of 3 W m^{-2} . Determine the value of R_2 that ensures that $V_1 = V_2$ for 3 W m^{-2} .
 - Is there any other particular light intensity where $V_1 = V_2$? Give reasons for your answer.
- Determine the resistance (R_x) of the LDR below at which the transistor just turns off. Assume the transistor just turns off when $V_{\text{be}} \approx 0.65 \text{ V}$, and that the transistor's base current (I_b) is much smaller than the current flowing in the R_x/R_2 voltage divider.

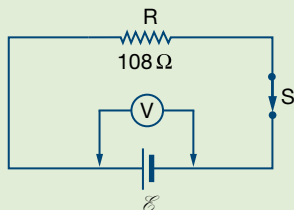


- If R_{LDR} increases to $10R_x$, what happens to V_{out} ?
- If R_{LDR} decreases to $R_x/10$ what happens to V_{out} ?

exam-style questions Electronics and photonics

For all questions, assume voltmeters and ammeters are ideal unless otherwise stated.

The following circuit applies to questions 1 and 2.



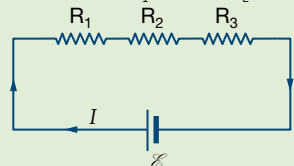
- 1 Assume that the battery has EMF $\mathcal{E} = 220 \text{ V}$ and internal resistance $R = 2.0 \Omega$ and that switch S is closed.

Calculate:

- the current I in the circuit
 - the reading on voltmeter V
 - the potential difference across R
 - the power output of the battery
 - the total power dissipated in the external resistor.
- 2 Calculate the terminal voltage of the battery if switch S is open.

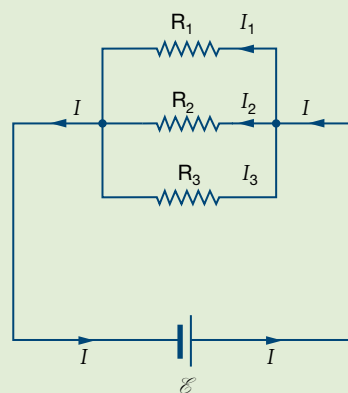
- 3 Consider the series circuit shown in the diagram.

Assume that $R_1 = 80 \Omega$, $R_2 = 10 \Omega$ and $R_3 = 8.0 \Omega$.

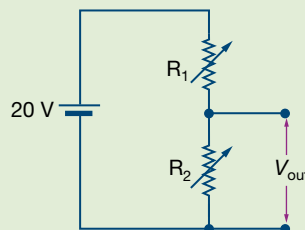


The battery has an EMF $\mathcal{E} = 100 \text{ V}$ and internal resistance $R = 2.0 \Omega$. Calculate:

- the total resistance R_t in the external circuit
 - the current I in the circuit
 - the potential difference across:
 - R_1
 - R_2
 - R_3
 - the terminal voltage V_t of the battery
 - the power output P of the battery.
- 4 Consider the following circuit where three resistors R_1 , R_2 and R_3 are connected in parallel. Assume that $R_1 = 100 \Omega$, $R_2 = 200 \Omega$ and $R_3 = 600 \Omega$. The battery has an EMF $\mathcal{E} = 120 \text{ V}$ and zero internal resistance.
- Calculate the total resistance R_t in the external circuit.
 - Calculate the line current I in the circuit.
 - Determine the branch currents I_1 , I_2 and I_3 .
 - What is the power output P of the battery?
 - Calculate the total power consumed by all the resistors in the external circuit.

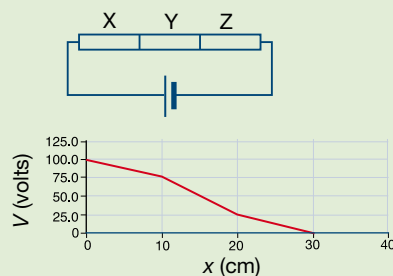


- 5 The circuit shown is a simple voltage divider. Complete the table that follows.



$R_1 [\Omega]$	$R_2 [\Omega]$	$V_{out} [\text{V}]$
1000		10
	1000	5.0
400	100	
900		2.0
2.0	3.0	
	100	4.0

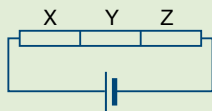
- 6 Three uniform sections of different types of resistance wire, X, Y and Z, each of length 10 m, are connected in series to an ideal battery as shown. The graph shows the electric potential V at any point along the composite wire as a function of the distance x from the positive terminal.



- Determine the value of the ratio R_X/R_Y .
- Determine the value of the ratio R_Y/R_Z .
- What is the electrical energy gained by an electron as it travels through section X of the composite wire? [$e = 1.60 \times 10^{-19} \text{ C}$]

- d What is the electrical energy gained by an electron as it travels through section Y of the composite wire?
- e Determine the EMF of the battery.

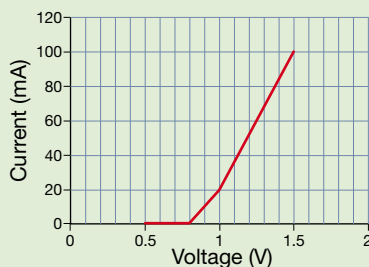
- 7 Three uniform sections of resistance wire, X, Y and Z, each of length 50 cm and composed of different materials, are connected in series to an ideal battery of EMF $\mathcal{E} = 550 \text{ V}$ as shown in the diagram.



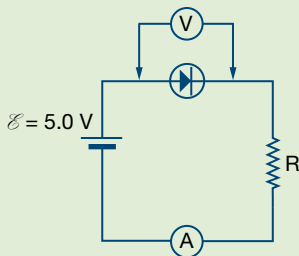
The respective resistances of each wire are $R_X = 100 \Omega$, $R_Y = 200 \Omega$, $R_Z = 250 \Omega$.

Draw the graph of electric potential [V] as a function of the distance x from the negative terminal of the battery.

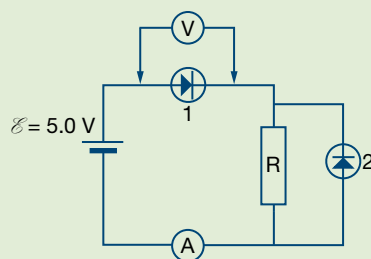
- 8 a What is a semiconductor?
- b Explain the difference between n-type silicon and p-type silicon semiconductor material.
- c What is a pn junction?
- 9 Explain the terms 'reverse bias' and 'forward bias' in reference to a pn junction.
- 10 A diode has an I - V graph as shown, and a switch-on voltage of 0.8 V . When the diode is connected in the circuit as shown, the voltmeter reading is 1.0 V .



- a Determine the current flowing through the diode.
- b How much power is being consumed by the diode?
- c What is the value of R ?
- d What is the total power consumption in the circuit?
- e What is the reading on ammeter A?



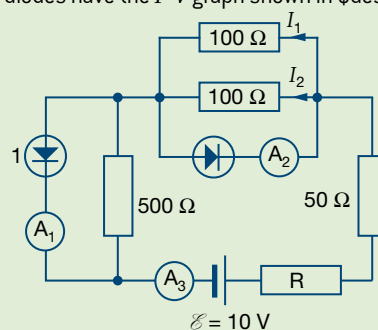
- 11 Two diodes have the same I - V graph as shown in question 10. When the diodes (1 and 2) are connected in the circuit as shown, the reading on the ammeter A is 100 mA .



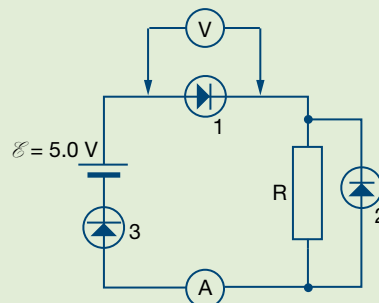
- a What is the reading on voltmeter V?
- b Calculate the power consumption of diode 1.
- c How much current is flowing through diode 2?
- d Calculate the voltage across diode 2.
- e Calculate the power consumption of diode 2.
- f Determine the value of R .
- g Calculate the power consumption of R .
- h Determine the power output of the battery.

The following information applies to questions 12–14.

In the circuit shown, the reading on A_1 is 20 mA . Assume that all diodes have the I - V graph shown in Question 10.

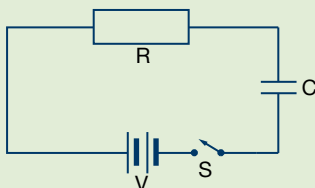


- 12 a What is the potential difference across diode 1?
- b Calculate the power consumption of diode 1.
- c How much current is flowing through ammeter A_2 ?
- 13 a How much current is flowing through ammeter A_3 ?
- b Calculate I_2 .
- c Calculate I_1 .
- 14 a Determine the value of R .
- b Calculate the power consumption of R .
- c Determine the power output of the battery.
- 15 Consider the circuit shown. The reading on voltmeter V is 1.0 V . The total power consumed by all diodes in the circuit $\Sigma P = 40 \text{ mW}$.



Calculate the value of R . Assume all diodes have the I - V graph shown in Question 10.

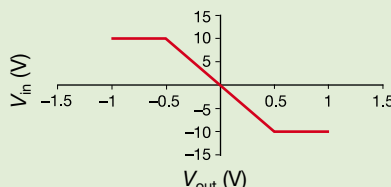
- 16 Explain the terms 'forward bias' and 'reverse bias' with respect to diodes.
- 17 What is meant by the term 'switch-on voltage' of a diode?
- 18 Consider a capacitor, $C = 1.0 \mu\text{F}$, connected in series with a resistor, $R = 1.0 \text{ k}\Omega$, and a battery, $V = 1.0 \text{ kV}$, as shown in the diagram. The switch is closed, resulting in the charging of the capacitor. Determine the final charge on the capacitor.



- 19 A capacitor, $C = 100 \text{ nF}$, is connected in series with a resistor R and a battery. The final charge on the capacitor is $50 \mu\text{C}$. What is the EMF of the battery?

The following information applies to questions 20 and 21.

The graph describes the characteristics (V_{out} vs. V_{in}) of a voltage amplifier. All voltages are in volts.

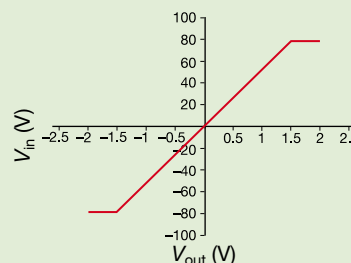


- 20 Explain why this amplifier is referred to as an inverting amplifier.
- 21 a What is the maximum peak signal voltage that can be amplified without distortion?
- b Calculate the voltage gain of the amplifier.
- 22 Complete the following table by calculating the range of output voltages produced by an amplifier with gain = -20 , for the following sinusoidal input voltages.

Peak input voltage	Output voltage range
100 mV	
0.25 V	
1.0 V	
20 mV	
0.80 V	

- 23 A particular amplifier has the following characteristics (V_{out} vs. V_{in}).
- a What type of amplifier is it? Explain.
- b What is the gain of the amplifier?

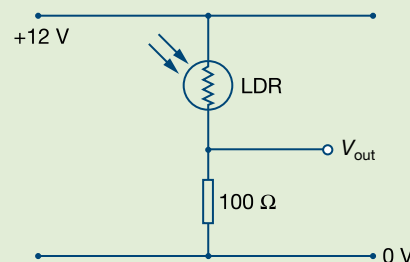
- c Determine the peak-to-peak output voltage when the peak-to-peak input voltage is 0.10 V



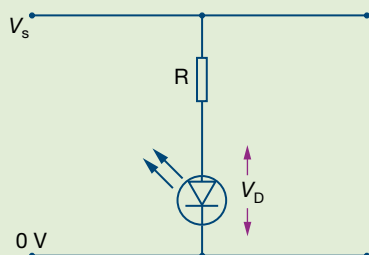
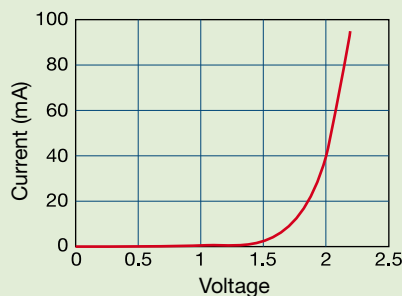
- 24 A student measures the resistance R of a thermistor for various temperatures T , and records the results in the table below:
- a What is a thermistor?
- b Plot a graph of R versus T and describe whether the relationship between resistance and temperature is linear.

$R \text{ (k}\Omega\text{)}$	$T \text{ (}^\circ\text{C)}$	$R \text{ (k}\Omega\text{)}$	$T \text{ (}^\circ\text{C)}$
20.0	2.5	2.5	30
15.0	5.0	2.0	40
10.0	10	1.8	45
5.0	20	1.0	50

- 25 In order to increase the overall gain, two or more transistor amplifiers may be connected together in a suitable way. Explain the role of capacitor coupling in such a situation.
- 26 In the simple LDR light detector circuit shown, V_{out} is to be used to activate an alarm when the ambient light reaches a certain level. At this particular light level, the resistance of the LDR is 200Ω . The alarm activates whenever V_{out} is above the trigger level.



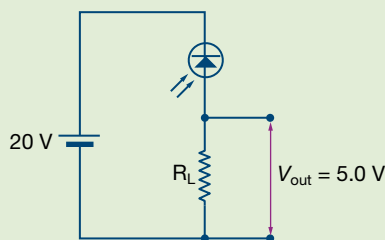
- a What is the value of V_{out} at which the alarm should activate?
- b Will the alarm activate when the light is above or below the particular level of concern? Explain your answer.
- c When it is very dark, what would you expect V_{out} to become?
- 27 The graph shows the characteristics of an LED which is to be used in the circuit shown. For optimum life and light efficiency the current through the LED should be 40 mA . At higher currents the LED will be brighter, but its life is shortened and it will burn out rapidly if the current exceeds 90 mA .



- With a power supply $[V_s]$ of 10 V, what is the optimum value for R ?
- What is the minimum possible value for R if the LED is not to burn out rapidly?
- If a resistor of $100\ \Omega$ is used for R , approximately how much current will flow in the LED? How would you describe the condition of the LED in this case?
- If the voltage across the LED $[V_D]$ is less than 1.7 V, it will be too dim to see. What is the minimum value of the supply voltage that could be used with the $100\ \Omega$ resistor if the LED is to be just visible?

- What is a photodiode and how does it work?
 - Explain the meaning of the term 'dark current' in relation to a photodiode.
 - Under what circumstances will the current through a photodiode circuit be zero?

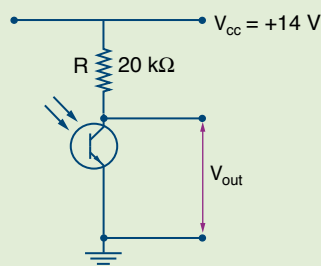
- A senior physics student designs a circuit containing a photodiode detector with active area $16\ \text{mm}^2$ as shown in the diagram. The specifications of the photodiode used in the circuit indicate that in photoconductive mode the device generates a photocurrent of $I = 10\ \mu\text{A}$ under illumination of $\phi = 2.0\ \text{W m}^{-2}$. The device is operating under a constant illumination of $\phi = 2.0\ \text{W m}^{-2}$.



- Determine the value of R_L that will produce $V_{\text{out}} = 5.0\ \text{V}$.
- What is the radiation power absorbed by the active area of the detector?
- Calculate the power dissipation in the diode.
- How much power is being dissipated in R_L ?
- Determine the power output of the battery.

- What is a phototransistor? Explain its operation.

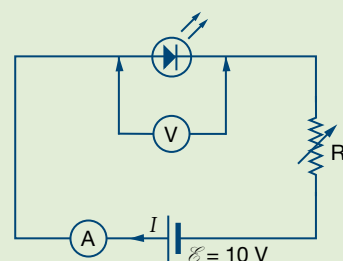
- A phototransistor has an active detection area $A = 2.25\ \text{mm}^2$. It is illuminated with green light of $\lambda = 536\ \text{nm}$ and intensity $\phi = 4.0\ \text{W m}^{-2}$. The phototransistor is connected into the circuit as shown in the diagram.



Assume: $h = 6.63 \times 10^{-34}\ \text{J s}$, $c = 3.00 \times 10^8\ \text{m s}^{-1}$, $e = 1.60 \times 10^{-19}\ \text{C}$.

- Assuming that all incident light is absorbed, calculate the optical power absorbed by the phototransistor.
- Determine the energy of each photon in the incident light beam.
- Calculate the number of photons that are absorbed each second by the phototransistor.
- Assuming that each absorbed photon generates one photoelectron into the base of the phototransistor, determine the base photocurrent, I_{pb} .
- Determine the photocurrent gain for the phototransistor when $V_{\text{out}} = 10\ \text{V}$.

- An LED constructed from gallium phosphide (GaP) is connected into the circuit shown. A student records the variation in current I as a function of the applied forward voltage V . The results are shown in the table.



V (V)	I (mA)
1.0	0
1.2	0
1.4	0
1.6	0.01
1.8	3.0
2.0	5.0
2.2	10

- a** What is a light-emitting diode and how does it produce light?
- b** Estimate the threshold or switch-on voltage for this LED.
- c** Determine the value of R when the current through the LED is 5.0 mA.
- d** What is the power consumption of the diode when $V = 2.20$ V?
- e** What is the power output of the battery when $V = 2.20$ V?
- 33** Which of the following devices is not a photonic transducer?
- A** LED
- B** LDR
- C** photodiode
- D** solar cell
- E** capacitor

Unit

area of study 3

3

Detailed studies

detailed studies

Chapters 6–8 are the detailed studies for Unit 3. You will undertake one detailed study in each unit.

Chapter 6 Einstein's special relativity
Chapter 7 Materials and their use in structures
Chapter 8 Further electronics

Einstein's special relativity

At the end of the 20th century, *Time* magazine chose Albert Einstein as its 'Person of the Century'. The image of the shaggy-haired physicist is instantly recognisable to most people 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 the Second World War, 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.

Galileo and Newton laid the foundations of the 'clockwork universe', a mechanical picture of the world which has underpinned most modern world views and philosophies. In fact, some would say that it has led to the excessive materialism which now threatens our planet. Einstein, along with others such as Bohr and Heisenberg, has brought us a much richer and more mysterious universe, one which challenges us to think beyond the mechanical, materialistic picture we so often take for granted. Perhaps this new way of thinking will eventually flow through to our social and cultural consciousness and help us to see our world and its people in a new way.

by the end of this chapter

you will have covered material from the study of Einstein's special relativity, 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 time dilation and length contraction equations
- an explanation of the results of the Michelson–Morley experiment
- a comparison of the relativistic and Newtonian models of motion
- some of the consequences and applications of special relativity.

outcome

On completion of this chapter, you should be able to use Einstein's theory of relativity to describe and explain relativistic motion and effects, and make comparisons with classical descriptions of motion.

6.1 Two principles Einstein did not want to give up

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 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 that 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.



Figure 6.1 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 the physicists's 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 Melbournians. (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 Melbournians 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.

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 motion, a state of rest simply being a special case of uniform motion. This was quite contrary to the beliefs of the Aristoteleans, 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 $+25 \text{ m s}^{-1}$, 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!

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. So 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's 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, Chloe, is watching this ball game from beside the track. Chloe 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 6.1

Anna and Ben 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 6.3. Anna, who is at the front of the carriage, throws the ball at 10 m s^{-1} . Ben catches it and throws it back at the same speed.

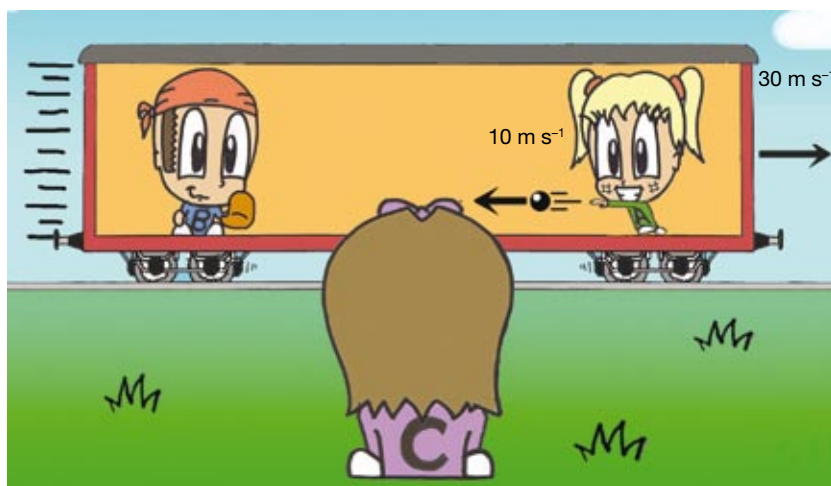


Figure 6.3 Anna has thrown the ball to Ben at 10 m s^{-1} , but how fast is the ball really moving?



Figure 6.2 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 $30\,000 \text{ m s}^{-1}$.

- Chloe, 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?
- Ben 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 Chloe's point of view?

Solution

We will choose the direction of the train's motion as positive and so Anna throws the ball with a velocity of -10 m s^{-1} in the train frame of reference.

- a** Chloe sees Anna throw the ball backwards relative to the train's motion. Its velocity will be:

$$30 + (-10) = 20 \text{ m s}^{-1}$$

When Ben throws it back, the ball will be moving at:

$$30 + 10 = 40 \text{ m s}^{-1} \text{ relative to Chloe}$$

- b** To determine the force, we need to know the ball's acceleration. The change of velocity, from Anna and Ben's point of view is:

$$+10 - (-10) = +20 \text{ m s}^{-1}$$

From Chloe's point of view the change of velocity was:

$$+40 - 20 = +20 \text{ m s}^{-1}$$

This change took 2 s and so in both frames of reference the acceleration was:

$$\frac{20}{2} = 10 \text{ m s}^{-2}$$

And the force Ben exerted was therefore:

$$F = ma$$

$$= 0.2 \times 10$$

$$= 2 \text{ N}$$

You were probably not surprised about the result in this example. Indeed it would be very strange if the force Ben 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, 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.



Figure 6.4 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.

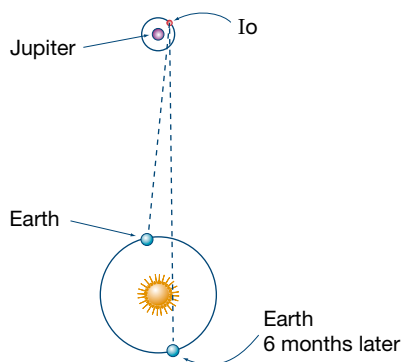


Figure 6.5 Ole Römer determined the time taken for the light from Jupiter's moons to cross the diameter of the Earth's orbit.

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 6.6), 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:

$$c = 2.997\,924\,58 \times 10^8 \text{ m s}^{-1}$$

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. So what type of wave was light, and what was the medium that carried it? 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.



Figure 6.6 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.

Physics file

The value of $c = 2.997\,924\,58 \times 10^8 \text{ m s}^{-1}$ 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 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.

Maxwell's equations

$$\oint \vec{E} \cdot d\vec{A} = q/\epsilon_0$$

$$\oint \vec{B} \cdot d\vec{A} = 0$$

$$\oint \vec{E} \cdot d\vec{s} = - \frac{d\Phi_B}{dt}$$

$$\oint \vec{B} \cdot d\vec{s} = \mu_0 \epsilon_0 \frac{d\Phi_E}{dt} + \mu_0 i$$

Figure 6.7 These 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!

Maxwell's conundrum

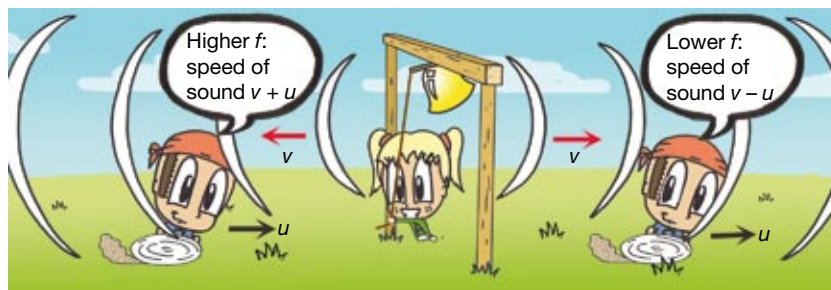
In 1864, 250 years after Galileo's major work, Maxwell introduced the second radical idea crucial to our story of relativity. Maxwell was a brilliant mathematician and theoretical physicist. He had taken Faraday's imaginative concept of electric and magnetic fields and worked with them to produce a mathematical description that encompassed all known electromagnetic phenomena. This description can be summed up in four famous equations referred to as Maxwell's equations (Figure 6.7).

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 $3 \times 10^8 \text{ m s}^{-1}$, 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.)

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 $3 \times 10^8 \text{ m s}^{-1}$ —something quite against all the known laws of physics, and particularly in complete contradiction to the principle of Galilean relativity.

Figure 6.8 As Ben runs towards Anna 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.



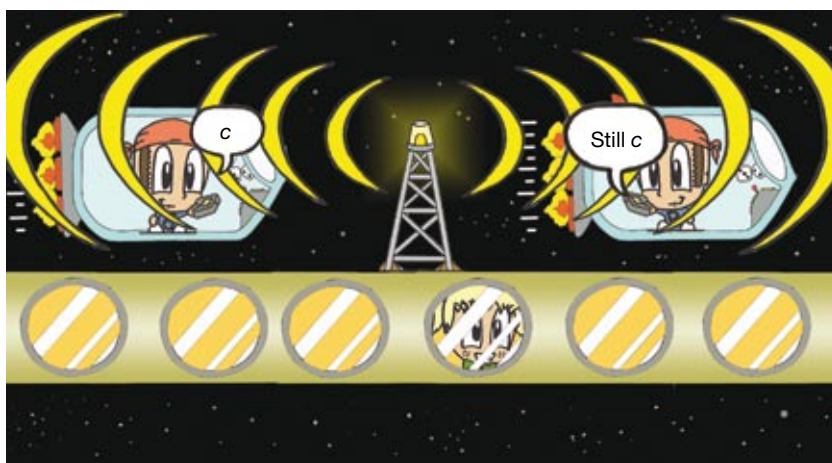


Figure 6.9 This time, as Ben speeds towards Anna's space station beacon light, he measures the apparent speed of light to be c . Maxwell's equations say that as Ben moves away from the space station, he will find that the speed is still exactly c .

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 may 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,

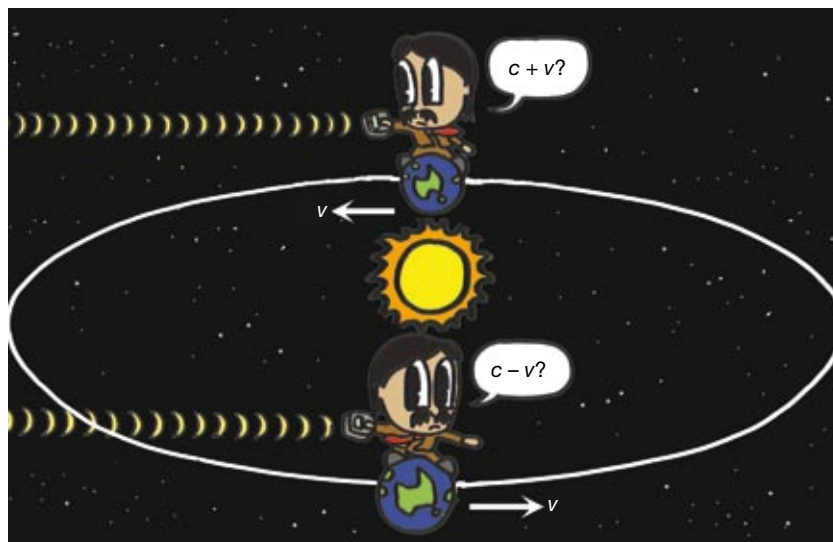


Figure 6.10 The basic principle of the Michelson–Morley experiment. If the aether is fixed relative to the Sun, and the light is travelling [at c relative to the aether] in the same direction as the Earth, the apparent speed should be less than c . Six months later, light travelling in the same direction should appear faster than c . The details were actually a little more complex.

the apparent speed should be faster. The differences would be tiny, less than 0.01%, but Michelson was confident that he could measure them.

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. (See the Physics in action, page 195.) 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.

Physics in action

The Michelson–Morley experiment

The principle of the Michelson–Morley experiment is shown in Figure 6.10, but in practice it was not quite so simple. It was not possible to directly measure the speed of light to the accuracy needed, so Michelson and Morley compared the speed of light on two paths perpendicular to each other. The comparison was achieved by reflecting light from a single source back and forth between two sets of mirrors mounted at right angles and fixed rigidly to a block of granite floating on a bath of mercury. (It was extremely important to ensure no movement between the mirrors as the apparatus was rotated.) After reflection, the light from the two beams was recombined and so interference occurred, giving a pattern of light and dark bands. The experimenters expected that the light travelling back and forth in the direction of the Earth's velocity would take a slightly longer time to cover the distance than light travelling at right angles to that direction. An analogy will help us to see why.

Although the analogy is not quite correct, one can imagine two boats motoring on a river flowing at 3 m s^{-1} . Both travel at the same speed, let's say 5 m s^{-1} , through the water. Imagine that one boat travels 1 km up the river and then 1 km back down the river. Due to the river current, its speed

upstream will only be 2 m s^{-1} and its speed downstream 8 m s^{-1} . The total time for the 1 km trip will therefore be:

$$\frac{1000}{2} + \frac{1000}{8} = 625 \text{ s}$$

The other boat travels 1 km and back directly across the river. Due to the river current, it has to point somewhat upstream and so its speed across the river will be slower than its speed through the water. Pythagoras's theorem shows that this boat will travel at 4 m s^{-1} each way. Thus the time taken will be:

$$\frac{1000}{4} = 500 \text{ s}$$

This is considerably faster than the other boat. This is a general result; the boat travelling upstream and downstream always takes a little more time than the boat travelling across the river.

Michelson and Morley therefore expected that when they rotated their whole apparatus through 90° , there would be a shift in the interference pattern due to the small difference in times for light to traverse the two paths. In fact, they found no shift in the pattern, indicating that the light was travelling along both paths at exactly the same speed.

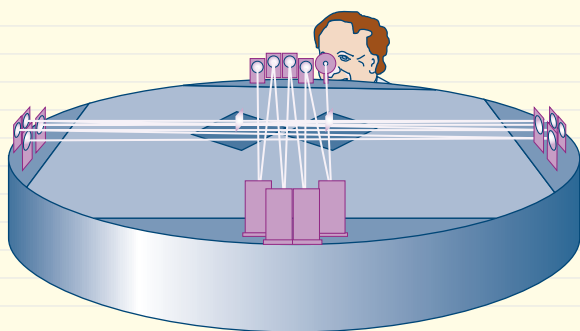


Figure 6.11 Michelson and Morley compared the time for light to travel on two perpendicular paths.

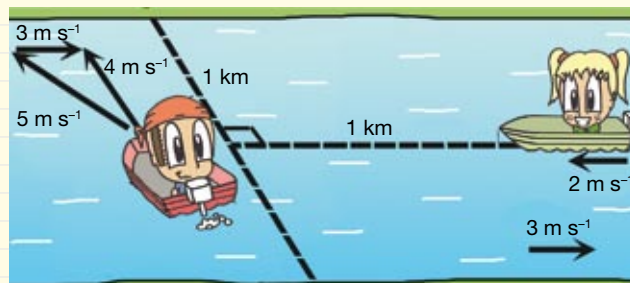


Figure 6.12 Anna's boat travels 1 km up and back down the river while Ben travels across the river perpendicular to the shore. Both boats travel at 5 m s^{-1} relative to the water. Ben has to head somewhat upstream and, as can be seen from the vector diagram, his speed across the river is 4 m s^{-1} .



6.1 summary

Two principles Einstein did not want to give up

- 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 $3.00 \times 10^8 \text{ m s}^{-1}$.
- Experiments such as Michelson and Morley's failed to detect the Earth's motion through the aether.



6.1 questions

Two principles Einstein did not want to give up

In the following questions, use 6400 km as the Earth's radius and 150 million km as the radius of the Earth's orbit around the Sun.

- 1 What was the key difference between Aristotle's ideas on motion and those of Galileo?
- 2 How did Galileo's law of inertia imply that there was no absolute frame of reference?
- 3
 - 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?
- 4 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?
- 5 If you walk up a moving escalator, does it require more muscle force than if you walk up a flight of stairs? Explain.
- 6 Apart from safety issues, why do airlines not serve drinks while the aircraft is taking off?
- 7 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?
- 8 If the speed of sound in air is 340 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 m s^{-1} .
 - b You are driving away from the stationary truck at 40 m s^{-1} .
 - c You are stationary and the fire truck is heading towards you at 20 m s^{-1} .
 - d You are driving at 30 m s^{-1} and about to overtake the fire truck, which is travelling at 20 m s^{-1} in the same direction.
- 9 In each of the situations in Question 8, is the frequency of the sound you hear the same, higher or lower than if you and the truck were at rest?
- 10 Why did the physicists of the late 19th century feel the need to invent the idea of the aether?
- 11 Some philosophers claim that Einstein introduced the idea that 'everything is relative'. Is this a reasonable claim?
- 12 Calculate and compare the respective times taken for Anna's and Ben's boats (Figure 6.12) to travel their 2 km courses for a river speed of 0 m s^{-1} , 1 m s^{-1} and 4 m s^{-1} .

6.2 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, *Gedanken* as it is in 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.

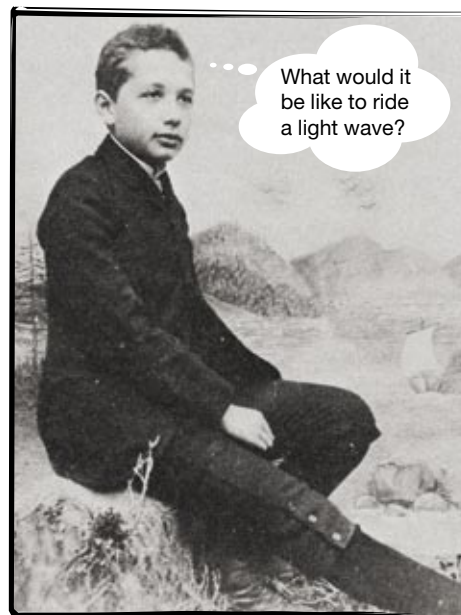


Figure 6.13 Einstein as a teenager.



Figure 6.14 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!

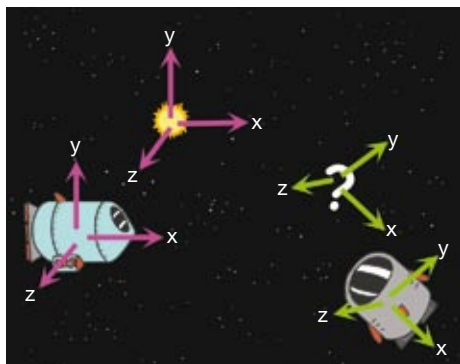


Figure 6.15 We can make measurements relative to stars and spaceships, but there is nothing in 'free space' to attach our coordinates to.

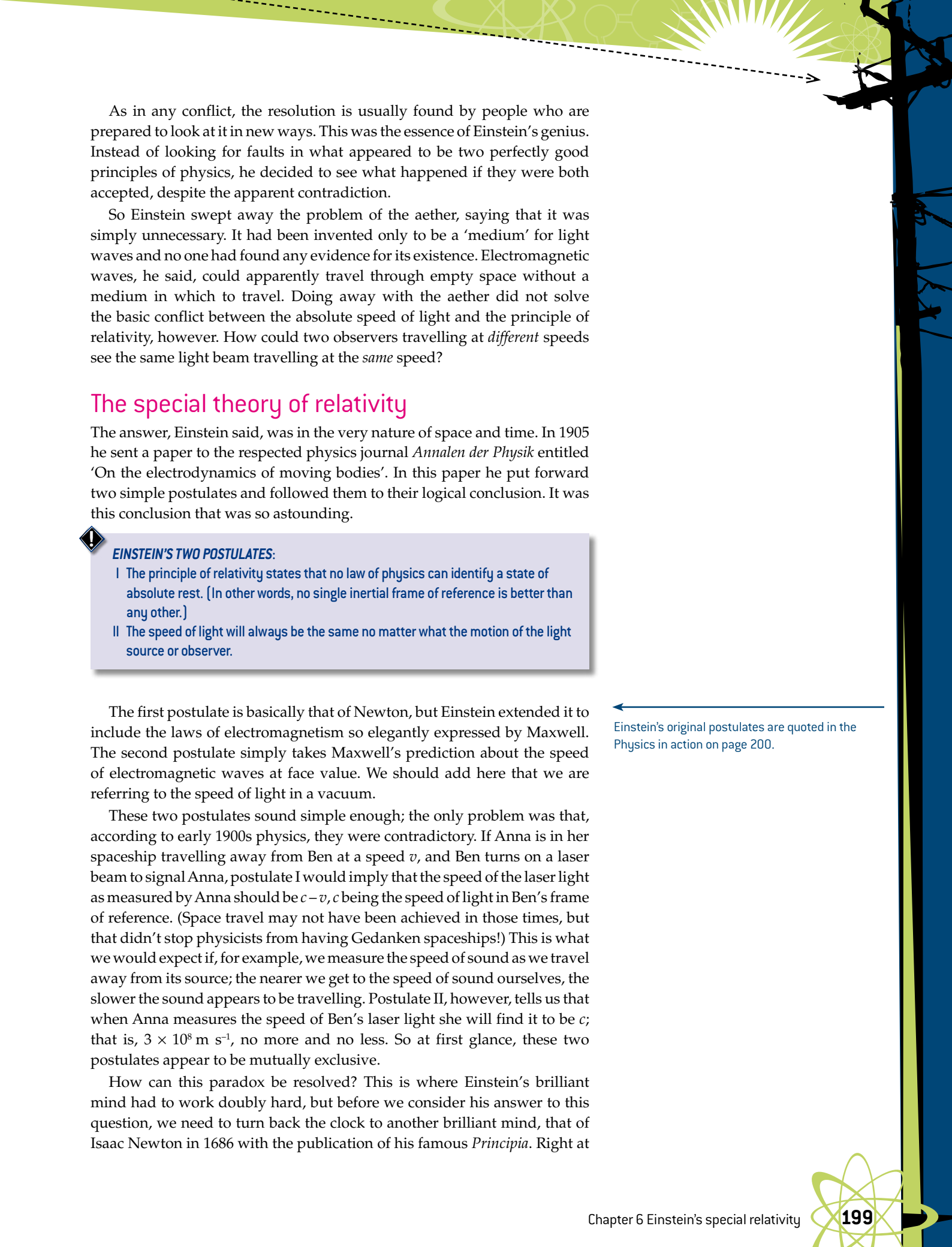
He expanded the Galilean principle to state that all inertial frames of reference must be equally valid, and that the laws of physics must apply equally in any frame of reference that is moving at a constant velocity. A consequence of this is that there is no physics experiment you can do, completely 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 get a feel for the reason 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, let's say, 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 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's take our veelo out into space a long way from our solar system.

In the distance we see a Vegarian spaceship (from the star system Vega) 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 Vega in the other craft. So if we use our veelo, what speed could it tell us about? And if the Vegarians 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.

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 Anna is in her spaceship travelling away from Ben at a speed v , and Ben turns on a laser beam to signal Anna, postulate I would imply that the speed of the laser light as measured by Anna should be $c - v$, c being the speed of light in Ben'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 Anna measures the speed of Ben's laser light she will find it to be c ; that is, $3 \times 10^8 \text{ m s}^{-1}$, 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

← Einstein's original postulates are quoted in the Physics in action on page 200.

Physics file

It should be pointed out here that even if the speed of light is the same for all observers, motion toward 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 6.8). The well-known red shift 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 to 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.

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.

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.

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

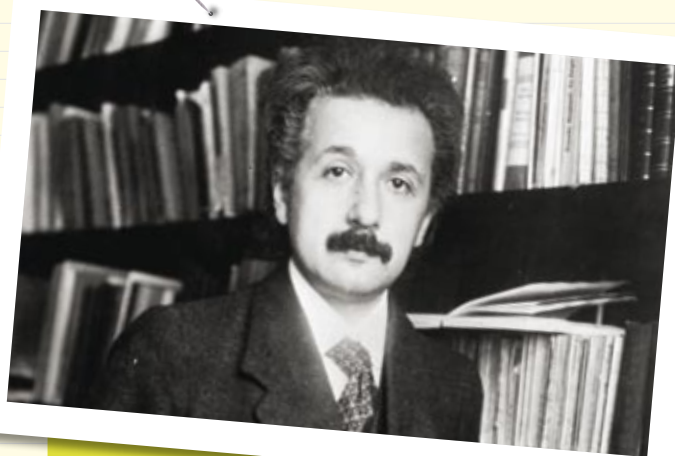


Figure 6.16 Einstein in 1905.

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*

Ed. John Stachel, 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 Anna and Ben are not convinced that the Earth is round; after all, it looks pretty flat. They decide to do an experiment to check. Anna heads north and Ben 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 Anna and Ben 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 6.17 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.

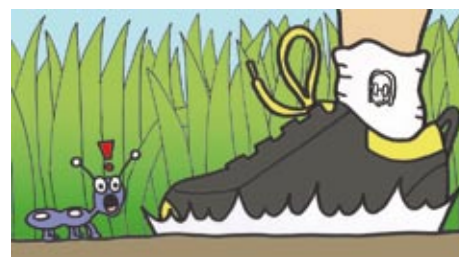


Figure 6.18 When the foot appears in the ant's 2D world, the ant has no idea where it came from.

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.



Figure 6.19 Anna and Ben are in the centre of the carriage watching for the reflections of the light flashes. Chloe is watching from the platform as the train speeds by.

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. Anna and Ben have boarded Einstein's train to help us with our experiments and Chloe is still outside on the platform. This train has a flashing light bulb set right in the centre of the carriage. Anna and Ben 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.

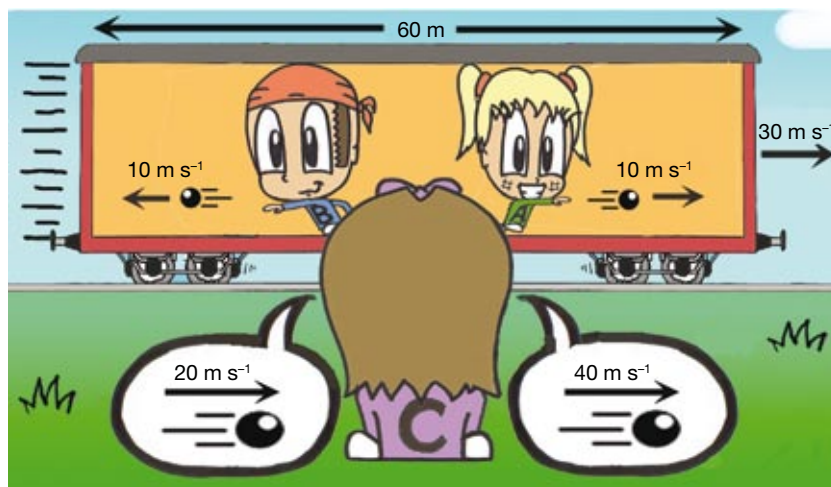
Outside, Chloe is watching the same flashes of light. Einstein's interest was in when Chloe saw the flashes hit the end walls. Now we realise that the light has to reflect from the walls and then travel to Chloe before she can 'see' it hit the wall, but we are going to make the simplifying assumption that she makes the appropriate calculations to determine when the light 'actually' hit the walls.

To appreciate Einstein's ideas, we need to contrast them with what we would normally expect. Consider the earlier example in which Anna and Ben were throwing balls back and forth in the train. It is important to appreciate that while our outside observer (Chloe) sees the various velocities involved differently, the times at which various events occur must be the same. This is illustrated in Worked example 6.2A.

Worked example 6.2A

The train carriage our Gedanken people are riding is 60 m long and travelling at 30 m s^{-1} . If Anna and Ben throw the balls at 10 m s^{-1} backward and forward from the centre, how long after being thrown do the balls reach the front and rear of the carriage from:

- within the train's frame of reference (Anna and Ben)?
- from the outside frame of reference (Chloe)?



Solution

- a** Anna and Ben see the balls take 3 s to travel the 30 m from the centre to the ends of the carriage.
- b** Chloe sees Anna's ball, thrown forward, travelling with the velocity of the train plus the 10 m s^{-1} that Anna gave it, i.e. at 40 m s^{-1} . But as the ball is moving forward, the train wall is also moving forward. We could set up a simultaneous equation pair to solve this problem, but we suspect that we already know that the answer will be 3 s and check that that is the case:
- In 3 s the train moves 90 m.
- In 3 s Anna's ball will travel $3 \times 40 = 120 \text{ m}$.
- This 120 m is in fact the 30 m length of the carriage plus the 90 m the train moved. The ball and train therefore 'meet' 120 m down the track. So 3 s is the correct solution, just as we thought.
- In a similar way, the ball that Ben threw backwards is actually travelling forward at a velocity of 20 m s^{-1} in Chloe's frame.
- In 3 s Ben's ball will travel $3 \times 20 = 60 \text{ m}$ (forward).
- In the same time, the rear wall of the train moved 90 m and will therefore have caught up to the ball (which started 30 m up the track).



PRACTICAL ACTIVITY 23

Relative motion in one dimension

If we had discussed a pulse of sound waves travelling from the centre of the train we would find exactly the same result: Chloe always agrees with Anna and Ben 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?

Einstein's second postulate tells us that *all* observers see light travel at the same speed. Anna, Ben and Chloe will all see the light travelling at $3 \times 10^8 \text{ m s}^{-1}$, 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!

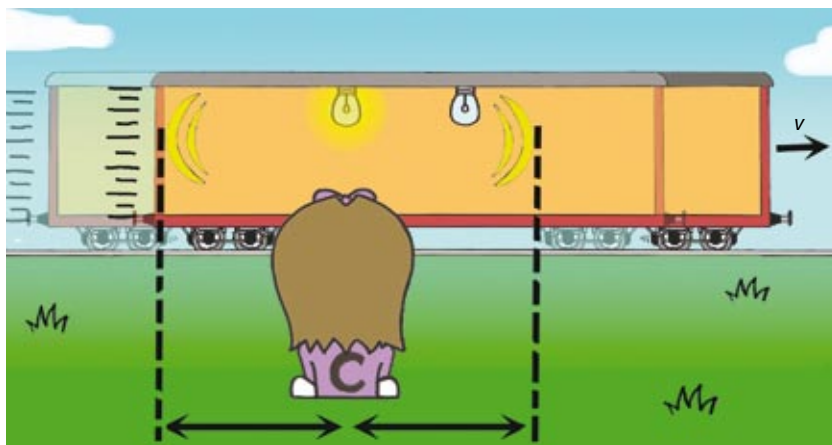


Figure 6.20 Chloe sees the flash hit the back end before the front end.

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 Chloe'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—Anna and Ben in the moving air, and Chloe in still air—would have agreed on the time difference. The strange thing about light is that the observers *don't* agree on the times!

If Chloe sees the light travelling at the same speed in the forward and back directions 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. Anna and Ben saw the light flashes reach the ends of the carriage at the *same* time, Chloe 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 Chloe 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 Anna and Ben's frame of reference to Chloe'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 Anna and Ben 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 Chloe'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

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 Anna and Ben's frame will always be different from that in Chloe'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.



Figure 6.21 The famous clock tower in Berne, Switzerland, where Einstein did a lot of thinking about time.



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, Anna and Ben saw the light flashes that went forward and back take the same time to reach the walls, but in Chloe'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.

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.

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.



6.2 summary

Einstein's crazy idea

- Einstein decided that Galileo's principle of relativity was so elegant it simply had to be true.
- 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.



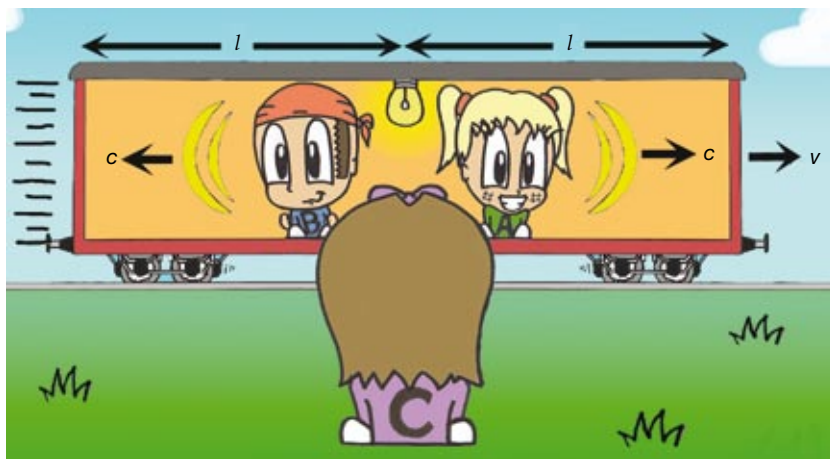
6.2 questions

Einstein's crazy idea

- 1 It is easy to stop light with a black surface, so why can't we see 'stationary light'?
- 2 Briefly explain why Einstein said that we could not catch up with light and see it 'stopped'.
- 3 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.
- 4
 - a Is the Earth's surface really an inertial frame of reference? Explain.
 - b Whereabouts on the Earth's surface are we closest to an inertial frame?
- 5
 - 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?
- 6 Two spaceships are travelling for a while with a constant relative velocity. Then one begins to accelerate. A passenger with a laser-based velocity measurer sees the relative velocity increase. How could this passenger tell whether it was his own or the other ship that began to accelerate?
- 7 Anna is at the front end of a train carriage moving at 10 m s^{-1} . She throws a ball back to Ben, who is 5 m away at the other end of the carriage. Ben catches it 0.2 s after it was thrown. Chloe is watching all this from the side of the track.
 - a At what velocity does Chloe see the thrown ball travelling?
 - b How far, in Chloe's frame of reference, did the ball move while in flight?
 - c How long was it in flight in Chloe's frame of reference?
- 8 Imagine that the speed of light has suddenly slowed down to only 50 m s^{-1} and this time Anna (still at the front of the 5 m train moving at 10 m s^{-1}) sends a flash of light towards Ben. (Ignore the look-back time effects for these questions.)
 - a From Anna's point of view, how long does it take the light flash to reach Ben?
 - b How fast was the light travelling in Ben's frame of reference?
 - c In Chloe's frame of reference, how far did the train travel in 0.1 s?
 - d How fast was the light travelling in Chloe's frame of reference?
 - e Approximately, when did Chloe see the light reach Ben?
- 9 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 Time is not what it seems

So far we have discussed in general terms the consequences of Einstein's two postulates when applied to a simple Gedanken situation—a moving train. Observers inside the train saw two simultaneous events while those outside saw the two events occur at different times. Certainly the differences are minute, well and truly unobservable in any practical train, but even for aircraft flying around the world the differences become, while very small, measurable. For subatomic particles in devices such as a synchrotron the relativistic differences become very large indeed and it is essential to use Einstein's equations in their design. In this section we return to Einstein's train with a simplified analysis of the situation which will enable us to get a feel for some of these equations.



Physics file

The introduction to Einstein's original paper is quite easy to read (see the Physics in action, page 200). After that it takes dedication and very good maths! The arguments we can present here are, of necessity, simplified versions. We have tried to indicate some of the simplifications along the way. Clearly, then, our discussion and mathematical illustrations are not meant in any way as a 'proof' of Einstein's theory, only as illustrations. Hopefully they are reasonably consistent with the real thing and will give you a feel for Einstein's incredible achievement. The real proof of relativity theory is that it works! In fact, all sorts of experiments have tested it and found it accurate to very high degrees of precision. Endeavours such as the global positioning system and the Australian Synchrotron, for example, are totally dependent on relativity theory.

Figure 6.22 Anna and Ben see the light take $\frac{2l}{c}$ s to reflect back to them. Chloe sees things differently.

This time we will compare the time taken for the light flash to go from the centre of the carriage to the ends and back to the centre again, a distance of $2l$. (If we want to compare times we must compare the times between two events that occur at the same place.) In the frame of reference of Anna and Ben in the train, the time for the light pulse to go from the centre to the end and back to them is simply given by:

$$T_A = \frac{2l}{c}$$

For Chloe, watching from the platform, the light takes a shorter time to reach the rear of the carriage and then a longer time to return to Ben. In Chloe's frame the light travels at c , but the train is travelling in the opposite direction at v and so we can say that, from this frame, the relative speed of light and train is $c + v$. Hence, the time taken for the light to reach the rear of the train is given by:

$$t_1 = \frac{l}{(c + v)}$$

(This is not to say that light is travelling at $c + v$ in any frame. It is travelling at c as measured in either frame.) After reflection, in Chloe's frame, the light travels back to the centre of the carriage, relative to the train's motion, at $c - v$. It therefore takes time $t_2 = \frac{l}{c - v}$ to return to the centre. The total time for the round trip is then:

$$\begin{aligned} T_c = t_1 + t_2 &= \frac{l}{c + v} + \frac{l}{c - v} \\ &= \frac{l(c - v + c + v)}{c^2 - v^2} \end{aligned}$$

Physics file

You may have noticed that we made an assumption here that seems reasonable, but which perhaps we should question! More about this later.

Physics file

The factor $1 - \frac{v^2}{c^2}$ is actually the *square* of a factor that occurs frequently in relativity:

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

Thus the ratio of the times

$$T_c = \frac{T_A}{\sqrt{1 - \frac{v^2}{c^2}}}$$
 can be expressed as:

$$T_c = \gamma^2 T_A$$

In fact, as we will see shortly, this is not the correct expression for this ratio. The correct one does not have γ^2 but simply γ .

Physics file

Chloe knows that she will need to correct for any look-back time effects. Whether she uses a telescope to look at the clock or picks up radio signals from it makes no difference to our discussion. Light and radio waves are just different forms of electromagnetic waves which all travel at c . When we talk of the 'speed of light', we are really talking of the 'speed of electromagnetic waves'. After having received the signal from the clock, Chloe will need to allow for the time taken for the signal to reach her from the distant spacecraft.

$$\begin{aligned} &= \frac{2lc}{c^2 - v^2} \\ &= \frac{2lc}{c^2(1 - \frac{v^2}{c^2})} = \frac{2l}{c(1 - \frac{v^2}{c^2})} \end{aligned}$$

Notice that whether we take the time for the light pulse to go to the back and return, or to the front and return, we get the same result.

You will notice (as the denominator must be less than c) that this time is greater than T_A . Chloe sees the time for the round trip by the light as longer than the time observed by Anna and Ben. As we deduced earlier, time appears (as observed from the outside) to go more slowly in a moving frame of reference. We can go one step further at this point and find the ratio of the two times. In this case, the length of the train will cancel out and we find:

$$T_c = \frac{T_A}{(1 - \frac{v^2}{c^2})}$$

Chloe's measured time appears to be longer than Anna's by a factor of $\frac{1}{(1 - \frac{v^2}{c^2})}$.

We can now get an indication of why we do not usually notice these time differences. A very fast train might travel at something over 100 m s^{-1} . This is less than a millionth of the speed of light. That is, $v/c = 10^{-6}$. This factor has to be squared and subtracted from 1. That makes the ratio different from 1 only in about the twelfth decimal place; relativistic effects are well outside the range of our normal experiences!

Before going further, it is important to point out that in arriving at the expression for the ratio of the two times, T_c/T_A , we made an assumption which seemed reasonable, but as we shall see shortly, in fact was not. We assumed that we could cancel the two lengths, l , in the equations. In other words, we assumed that the length, l , as measured by Anna and Ben was the same as that measured by Chloe. But was this actually a good assumption? After all, if time is behaving strangely, can we expect space (and therefore lengths) to be any better behaved? We will return to this question.

The light clock

If we really want to be sure that time appears to be 'flowing' more slowly in the moving train we need to watch a clock in the train to see if it is indeed going slow. As they have found that the effects of the motion on the time in the train are minuscule, Anna and Ben have now upgraded to a Gedanken spaceship in which they can travel at speeds a sizeable fraction of the speed of light. Chloe is going to watch from a space station, which we will regard as our stationary frame of reference. Anna and Ben have taken along a clock which Chloe can read from a distance.

Like any clock, this clock is governed by a regular oscillation, the period of which is a *tick*. Whether the oscillator is a pendulum, a rotating spring wheel or electrical vibrations in a quartz crystal, all clocks work by counting the number of ticks and converting that information into the appropriate number of hours, minutes and seconds.

Anna's clock has a very simple oscillator. It is just a light pulse that bounces back and forth between two mirrors. Some clever electronics convert the number of bounces, or ticks, into the time. (These clocks are only available

in Gedanken shops, by the way, and because of the constancy of the speed of light they are guaranteed to be precisely accurate!) Chloe has an identical clock in her own space station with which she can compare the tick rate of Anna's clock.

The advantage of this clock is that because the mechanism is so simple (to think about, not to construct!) we can predict how the motion will affect it. All we need is a little Pythagoras and some algebra. The clock has been set up in the spaceship so that the light pulses oscillate up and down a distance d at right angles to the direction of travel. As the ship speeds along, the light will trace out a zig-zag path as we see in Figure 6.24. We only need to consider one of the zigs, all the other zigs as well as the zags will have the same geometry. We will call the clock tick time the time for the light pulse to do one zig (or zag). Anna and Ben see a tick time (in their frame of reference) of T_A but Chloe, from her frame of reference will see, we suspect, a different time T_C . We want to find the relationship between these two times.

Anna and Ben see the light pulse travel the distance from one mirror to the other in time T_A and so they believe the distance, d , to be $c \times T_A$. On the other hand, Chloe sees the light travel a longer path (shown dotted in Figure 6.24). We need to find an expression relating Chloe's 'tick time' with the speeds and distances involved.

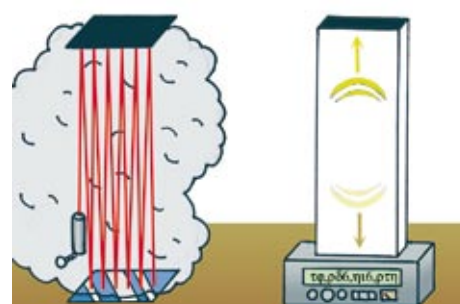
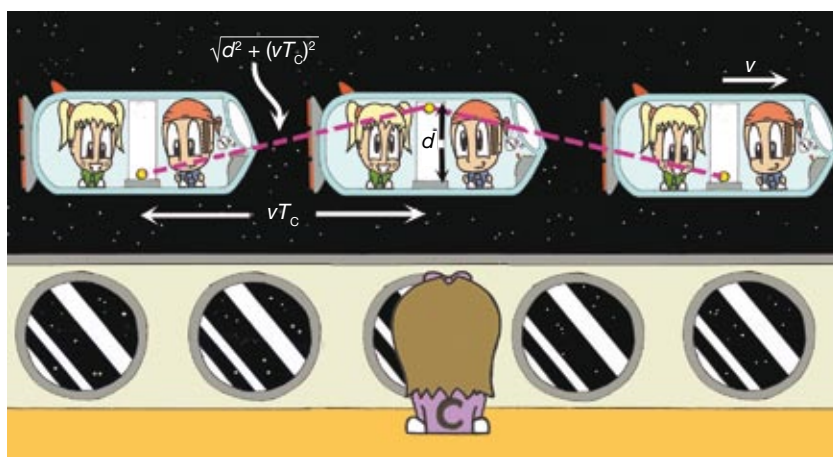


Figure 6.23 The concept of a 'light clock' as a beam of light pulses reflected between mirrors. The ideal 'Gedanken' light clock ticks each time the light pulse bounces at a mirror.

Figure 6.24 The light clock as seen by Chloe. The light pulses have to travel a greater distance between reflections and therefore the clock ticks more slowly according to Chloe.

Chloe sees the ship moving with a speed v , and so in one tick it will travel a distance equal to this velocity multiplied by the tick time, vT_C . In one tick the light pulse in the clock also travels the length d of the clock, making the combined distance, according to Pythagoras's theorem:

$$\sqrt{d^2 + v^2 T_C^2}$$

As always, the light pulse covers this distance at a speed of c , so we can write:

$$\sqrt{d^2 + v^2 T_C^2} = cT_C$$

If we square that expression and rearrange it, it becomes:

$$d^2 = (c^2 - v^2)T_C^2$$

From Anna and Ben's point of view (where $v = 0$), $d^2 = c^2 T_A^2$.

Note that the strange predictions of relativity result from the fact that we have used the same ' c ' in both these equations—something we would never do in classical physics, but something Einstein insists we must!

Physics file

It is worth pointing out that as Anna and Ben speed away from Chloe, the distance for the light from their clock to travel back to Chloe is increasing. This will result in Chloe seeing the clock apparently getting slower as each minute the time taken for the light to reach Chloe becomes slightly greater. This is *not* the reason for the time dilation! Chloe needs to correct for this effect as she makes her measurements. Notice that as the ship is coming toward her the opposite happens, the clock would appear to run fast. Chloe also has to correct for this effect, and when she does she finds again that their time is running slow. Mathematically, we can see that the time dilation results from the strange behaviour of light. As light travels on the diagonal zig-zag path it does so at speed c , not at a faster speed resulting from the additional component of the ship's motion as, for example, a boat zig-zagging across a river would as it is carried along by the current.

Physics file

You will probably remember that the factor $\frac{1}{1 - \frac{v^2}{c^2}}$ appeared earlier as the ratio of the travel time for light flashes in Anna and Ben's train as seen from Chloe's frame compared with theirs. This of course is γ^2 . There is a good reason for this apparent discrepancy and it will be revealed in section 6.4. In the meantime you might like to try to guess the explanation!

Table 6.1 The value of the Lorentz factor at various speeds

v/c [%]	γ
1	1.00005
10	1.0005
50	1.155
86.6	2.000
90	2.29
99	7.09
99.9	22.4

Now, although relativity warns us to be very careful when moving between frames of reference, hopefully the height of the clock, d , which is at right angles to the direction of travel (unlike the length of the train earlier), will be seen as the same by both sets of observers. Hence, we can equate the right-hand sides of these last two equations. This enables us to find the ratio of the times as seen in the two frames. A little more algebra produces:

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

Time dilation

This is Einstein's famous 'time dilation' equation. It relates the time elapsed as seen from a stationary frame of reference (Chloe's in our case) to that elapsed in a moving frame (Anna and Ben). The factor by which the time varies occurs frequently in relativity and is often abbreviated to γ (gamma):

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

So the time dilation equation becomes $T_C = \gamma T_A$. A quick inspection of the expression for γ will convince you that, as v^2/c^2 is almost zero for normal speeds, γ is normally very close to 1 (see Table 6.1). So for ordinary speeds, Chloe sees the tick time of the light clock on the spaceship to be the same as her own. However, when the spaceship has boosted to hyperdrive and is travelling at a (constant) speed of 99% of the speed of light, Chloe will measure a tick time for Anna and Ben's clock of seven times that of her own identical light clock. Time for them, Chloe says, seems to have slowed down to one-seventh of normal! Remember that this is Chloe's perception of time for Anna and Ben, not theirs. To Anna and Ben, time seems to be going at its normal rate. We can generalise this result.



EINSTEIN'S TIME DILATION EQUATION states that:

$$t = \gamma t_0$$

where t = the time as seen from the stationary frame
 t_0 = the time in the moving frame

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

As the denominator of the expression for gamma must be equal to or less than 1, gamma is always equal to or *greater than* 1 (although extremely close to 1 at any normal speed). Thus, time in the moving frame, as seen from the stationary frame, will always appear to elapse more slowly ($t \geq t_0$). The physicist H. A. Lorentz first introduced the factor γ in an attempt to explain the results of the Michelson–Morley experiment, so it is often known as the *Lorentz factor*.

It is important to realise that Anna and Ben do not perceive their time slowing down. To them, their clock keeps ticking away at the usual rate and their heartbeats take the same number of ticks. It is Chloe's measurement of Anna and Ben's time that goes slow. What might take her a few minutes would seem to take Anna or Ben much longer. It is as though they are moving in slow motion. However, we know that a slow-motion movie or an action replay are a type of illusion. On the other hand, Chloe's slow-motion view of Anna and Ben is no illusion.

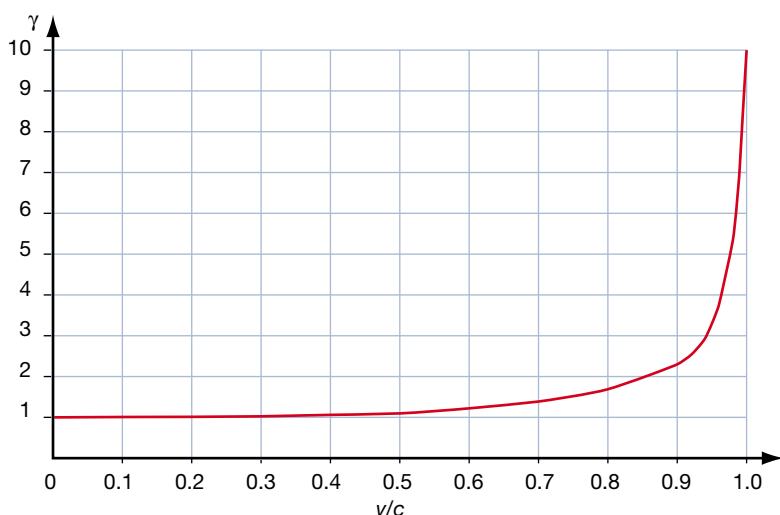


Figure 6.25 The graph of the Lorentz factor versus v/c .

Einstein's twins

We have been discussing the situation from Chloe's point of view, not Anna and Ben's. However, as all inertial frames of reference are equivalent, they are quite right to say that it is they who are at rest and it is Chloe's space station that is zooming along near the speed of light. This indeed is what the principle of relativity (Einstein's first postulate) is all about. So we could turn the whole argument around and have Anna and Ben watch the light clock in Chloe's space station. In which case we would find the reverse of our previous result—Anna and Ben would think that time has slowed down for Chloe! So who is right?

The answer is that they are both right. The whole point of relativity is that we can only measure quantities relative to some particular frame of reference, not in any absolute sense. Certainly Anna and Ben see Chloe as though in slow motion and Chloe sees them in slow motion. Remember, however, that there is no cosmic clock ticking away the absolute right time, and by which somehow we can discover the truth of the matter.

Nevertheless, many of us will feel that this situation is rather unsatisfactory. After all, what if Anna and Ben turn around and come back to meet Chloe on the space station? In that case we should be able to find out who has 'really' aged less. Indeed we can, and the answer is that it was Anna and Ben who aged more slowly! If all three were the same age at the start of these adventures, Chloe will find that she is older than the other two when they return. But surely Anna and Ben saw Chloe ageing more slowly?

This strange situation is often referred to as *Einstein's space twins paradox* because Einstein described one of a pair of twins setting off on a long space journey, only to find that when he returned his twin had died of old age long ago. While both observers in relative motion do see the other ageing more slowly, the key to this apparent paradox is that one observer has had to accelerate in order to get back to the other. It is the *acceleration* that makes all the difference!



Figure 6.26 As Chloe watches Anna and Ben play space squash, the ball seems to be moving much more slowly than in her own game.

Physics file

Einstein's general theory of relativity says that, in principle, being in a gravitational field is equivalent to being in an accelerated frame of reference. It also shows that in a strong gravitational field, and therefore in an accelerated frame of reference, time slows down. It is worth noting that for our space venturers to reach a speed close to c , they would need to accelerate at around g (9.8 m s^{-2}) for about a year! It is this acceleration in a non-inertial frame of reference that makes all the difference between their experience and Chloe's back in her inertial frame of reference. So will we live longer in an accelerated frame of reference? No! Remember that it is not our perception of time that changes, it is that of an outside observer in relative motion.

Here we need to point out that Einstein's 1905 theory of relativity deals only with frames of reference in *constant* relative motion (i.e. inertial frames of reference). For this reason it is called the *special theory of relativity*. Special relativity does not deal with accelerated situations. Ten years later, Einstein put forward the *general theory of relativity*, which does deal with situations in which acceleration occurs (i.e. with non-inertial frames of reference). As part of this theory, he showed that in an accelerated frame of reference, time also slows down.

Returning to our space venturers, in order for them to compare ages they need to meet again in the same place and time. This requires that at least one set of travellers undergoes acceleration as they slow down, turn around and head back to the others. Now the situation is no longer symmetrical; all agree that Anna and Ben underwent considerable acceleration in order to return to Chloe's space station, which has not accelerated at all. As Chloe watched from her inertial frame of reference, the theory of general relativity tells us that her view of Anna and Ben (ageing slowly) is more valid. As they accelerated, Anna and Ben's time slows relative to Chloe's and so they see Chloe's time speed up and more than make up for the apparent slowing down that occurred while they were in constant relative motion. It is as though the acceleration crystallises the time difference into place. As a result, they will have seen Chloe age more than they did, and when they all meet back at the space station that is indeed what they find. Although it is often called a paradox, there is actually nothing impossible or logically inconsistent about this story. Impracticable, certainly (at least at present), but not physically impossible.

Einstein himself pointed out that, due to the Earth's rotation (and thus acceleration), a clock on the Earth's equator would run a little more slowly than one at the poles. This has now been found to be the case. In fact, in 1971 accurate atomic clocks were flown around the world on commercial flights. When compared with those left behind, the difference of about a quarter of a microsecond was just what Einstein predicted! Nowadays, with many spacecraft in orbit, the theory has been well and truly tested many times. Indeed, global positioning systems (GPS) must take the relativistic corrections into account to ensure their accuracy.

Physics file

Recently there has been publicity given to research which has suggested that the speed of light is slowing down. Some have even suggested that Einstein's relativity itself is under threat! The research, based on analysis of light from very distant quasars, actually suggests that there have been very small changes in what is called the *fine structure constant*, which is made up of three more basic constants: the speed of light, the charge on an electron, and Planck's constant. Professor Paul Davies and others have suggested that if the evidence is correct, then it is probably the speed of light that is changing. If proved correct, no doubt this new data will modify some aspects of relativity, but to suggest that it will overturn relativity is a wild exaggeration. After all, while relativity certainly had an impact on Newtonian mechanics, it hardly overturned it, we still use Newton's laws for most of our mechanics. But the lesson from Einstein is that we must keep an open mind!

Worked example 6.3A

Imagine that one of a pair of twins takes off on a long space journey to Vega, 25 light-years away, at a speed, relative to Earth, of 99.5% of c ($\gamma = 10$). Once there he decides he doesn't like the Vegans, so turns around and comes straight back at the same speed.

- How long, in Earth's frame, does it take for the traveller to reach Vega?
- As seen by the Earth twin, how long does the trip take the traveller?
- How long does it take the traveller in his reference frame?
- Assuming a negligible turnaround time, how long did the trip take in the Earth's frame of reference?
- How long did the trip take the traveller?

Solution

- At $0.995c$, a trip of 25 light-years will take just over 25 years (25.1 years).
- b, c** Time for the traveller will seem to go at one-tenth ($1/\gamma$) the rate of Earth time, so when he gets there, his clocks (or calendar) will say that it took only 2.5 years. This is the time as seen in the traveller's frame from the Earth's frame, as well as in his own frame.

- d In the Earth's frame the trip took $2 \times 25.1 = 50.2$ years (plus turnaround time).
- e The traveller perceived that it took just 5 years (plus turnaround). So he returns to find his twin sister has aged over 50 years while he has only aged 5 years!

You may realise that there appears to be a problem here. How did the traveller manage to cover a distance of 50 light-years in 5 years? Did he really go faster than light? The answer is, of course, no! All will be explained in the next section and Worked example 6.4A.

Relatively intuitive

It is easy enough to see that velocity is a relative concept. You have probably had the experience of sitting in a stationary train that you thought started to move when it is actually the train on the next track moving the other way. It is not so easy to be convinced that time is relative; for example, we are quite sure that the watches worn by the people in the other train tell the same time as ours does. Because we don't normally experience relative time, we find it very hard to be really convinced about it.

Quite a bit of physics is intuitive. When we discuss momentum, for example, we can relate it to our experiences on the sports field, or in trying to control unwieldy skis on a slippery slope. Many concepts in physics are not very intuitive, however. It took over 2000 years of creative and imaginative thinking for humankind to arrive at the law of inertia, and even now many people still don't really believe it 'in their bones'. It is not surprising, then, that it is a little difficult to come to grips with the idea of relative time. The only way to do it really is to look at the evidence and follow the arguments. It also helps to know that now, just over 100 years after it was proposed, the experimental evidence has well and truly supported Einstein's original 'crazy idea'. Time *is* relative!

Having found that time is not the absolute and immutable quantity we thought it was, we may well ask what other quantities are not as they seem. The answer is that almost all quantities are affected by relativity. In fact, almost the only quantity not affected by relative motion is the speed of light!



6.3 summary

Time is not what it seems

- 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.
- Time in a moving frame seems to flow more slowly:

$$t = \gamma t_0$$

where t_0 is the time in the moving frame and γ is the Lorentz factor:

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

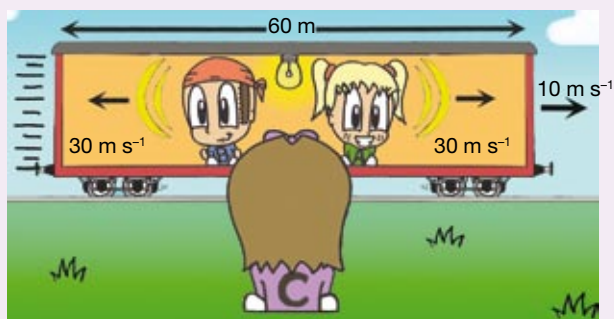
- Even at 10% of c , the Lorentz factor only makes about 0.5% difference.
- Observers in relative motion both see each other ageing more slowly, but if one accelerates in order to return to meet the other, that observer will have aged less than the other.



6.3 questions

Time is not what it seems

The following information applies to questions 1–7. The train carriage in the diagram is equipped with the famous flashing light in the centre. Anna and Ben, in the train, time the flashes as they are emitted, reflect off the end walls and return to Anna and Ben. Chloe is doing the same from the platform. (Chloe corrects her times for any look-back effects.) To make the numbers manageable we will slow light down to just 30 m s^{-1} . The train is moving at 10 m s^{-1} and is 60 m long.



- 1 According to Anna and Ben, how long does the light take to travel to the ends of the carriage and back to them?

In questions 2–6 consider the situation from Chloe's point of view.

- 2 The flash that travels to the back of the carriage will meet the wall early because the train is moving towards it.
 - a Chloe sees light travelling at 30 m s^{-1} in her frame, but at what speed *relative to the train* does she see the backward-directed flash moving?
 - b How long will it take this flash to reach the back wall?
 - c Use similar reasoning to find the time for the forward flash to reach the front wall.
- 3 a Continue the reasoning in Question 2 in order to find:
 - i the time, as seen by Chloe, for the flashes to return to Anna and Ben
 - ii the total time for the round trips for the forward and backward flashes.
- b How does this time compare with the time Anna and Ben measured for the same events?
- 4 Which events in these questions occurred simultaneously for both sets of observers and which were not simultaneous?

- 5 Had Anna and Ben been throwing balls or sending sound pulses, how would your answers to questions 1–4 have been different? Why would this have been the case?
- 6 In Question 2 we saw that Chloe 'saw' the light travelling at 40 m s^{-1} relative to the train. This is faster than the assumed speed of light (30 m s^{-1}). Does this contravene Einstein's second postulate? Explain.
- 7 Throughout these questions we have made an assumption which seems perfectly reasonable, but which Einstein would say we have to question. Can you think what that assumption might be?
- 8 Anna's Gedanken light clock has a height of 1 m between the mirrors, and relative to Chloe her spaceship is travelling at 90% of the speed of light ($c = 3.0 \times 10^8 \text{ m s}^{-1}$). One tick is the time for light to go from one mirror to the other.
 - a How far does the light flash travel in Anna's frame of reference in one tick, t_A ?
 - b So what is the tick time, t_A , for the clock in Anna's frame?

Now we know that, because the light takes a zig-zag path in her frame, Chloe sees the clock ticking at a slower rate, t_C . (Note also that in one tick Chloe sees the ship travel a distance $0.9ct_C$.)

 - c In terms of c and t_C what is the length of the zig path that the flash travels in one tick in Chloe's frame?
 - d Because the height of the light clock is 1 m we can now find (with some help from Pythagoras) the value of t_C . What is the tick time of the clock in Chloe's frame?
 - e What is the ratio of Chloe's tick to Anna's tick. How does that compare to the value of γ as found in Table 6.1?
- 9 If we repeat Question 8 but have Anna's ship travelling at $0.1c$, which answers would be different, and what would they be?
- 10 If Anna saw Ben fly by at $0.5c$, how long, in her frame, would it take Ben's clock to tick 1 second?
- 11 Briefly explain why Einstein said that a clock at the Earth's equator should run slightly slower than one at the Earth's poles. Why do we not find this to be a problem?

6.4 Time and space

Speed is a relationship between space and time. Einstein said the speed of light is the same for all observers. Observers in one frame of reference will see that time flows at a slower rate in another, moving, frame of reference. If the time is different, but the speed is the same, does that mean that distance, space itself, is different? In a word, yes. Einstein's special theory of relativity is about the strange relationship between light, time and space as seen from inertial frames of reference; that is, frames moving at constant relative velocities.

We already have a clue to the fact that lengths depend on who is measuring them. Earlier we found an expression for the apparent time ratio between Chloe's frame of reference and Anna and Ben's train carriage frame simply by working out the time for light to go from the centre of the carriage to the end and back. The expression was $T_C = T_A / (1 - v^2/c^2)$. This differs from Einstein's time dilation equation in that there is no square root sign in the factor $1/(1 - v^2/c^2)$. We could therefore write this version as $T_C = \gamma^2 T_A$. So why the difference? Where has the extra γ come from? The answer was hinted at when we derived the earlier expression. You will recall that in obtaining that result we cancelled out the length of the train, l , assuming that both sets of observers would see it as the same, normally a fairly reasonable assumption. By now, however, you will probably realise that we need to be very careful about assumptions we make concerning times and distances when high relative speeds are involved!

The light clock analysis is quite correct in its comparison of relative time. The light clock was used as it only depends on light, not some complicated mechanical arrangement which may well include other factors that are altered by relative motion. We did, however, make one other assumption in our clock analysis. It was that both Anna and Chloe would agree on the distance, d , between the mirrors. This enabled us to equate the two expressions for d in order to find the time ratio. Now the clock 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*.

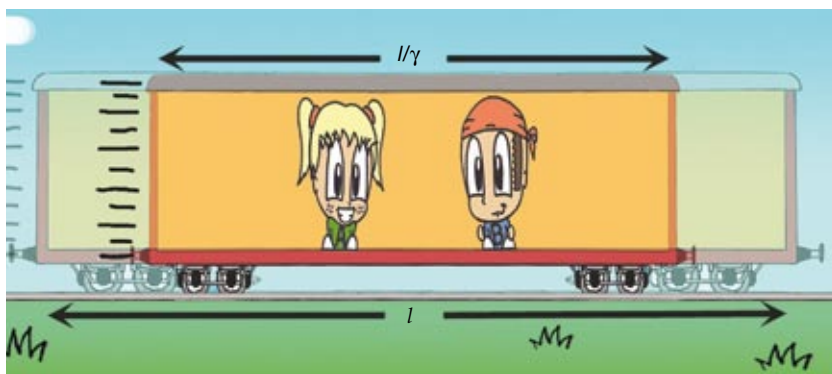


Figure 6.27 Einstein showed that a moving object is foreshortened by the γ factor. As a result, our earlier calculations for the time ratio, T_C/T_A , overestimated it by an extra γ .

Physics file

Length contraction is often referred to as *Lorentz contraction*. H. A. Lorentz first suggested it as a way of explaining the strange result of the Michelson–Morley experiment. He realised that if the apparatus itself shrank a little in the direction of its travel through the aether, this could explain the null result. He went on to calculate that the required contraction would be:

$$l = l_0 \sqrt{1 - \frac{v^2}{c^2}}$$

or, as we have expressed it:

$$l = \frac{l_0}{\gamma}$$

where γ is now known as the *Lorentz factor*.

The difference between his suggestion and Einstein's was that while Lorentz said that the apparatus physically shrank a little, Einstein said that it was not the apparatus, but space, along with the apparatus, that had contracted.

In the 'incorrect' expression for the time difference, Chloe saw time slow down even more (the extra γ) than she should have according to Einstein's time dilation equation. The reason for this was that the carriage actually appeared to be *shorter* than we assumed it was! In fact, if we redo the algebra with the l as seen by Chloe replaced by l_0/γ (where l_0 is the length measured by Anna), the extra γ cancels out and we end up with the correct expression: $T_C = \gamma T_A$. 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 Chloe would see the length l_0 , as measured by Anna and Ben, contracted by a factor γ .



The **LENGTH CONTRACTION EQUATION** is given by:

$$l = l_0 \sqrt{1 - \frac{v^2}{c^2}} \text{ or}$$

$$l = \frac{l_0}{\gamma}$$

Proper time and length

The time, t_0 , and the length, l_0 , are referred to as the *proper* time and length. They are the quantities measured by the observer who is in the same frame of reference. The proper time is the *time between two events which occur at the same point in space*. When the light bulb in the train flashes and Anna measures the time for the flash to return to her she has measured the proper time. Chloe does not measure the proper time because when the flash returned to Anna, Anna was not in the same place (in Chloe's frame).

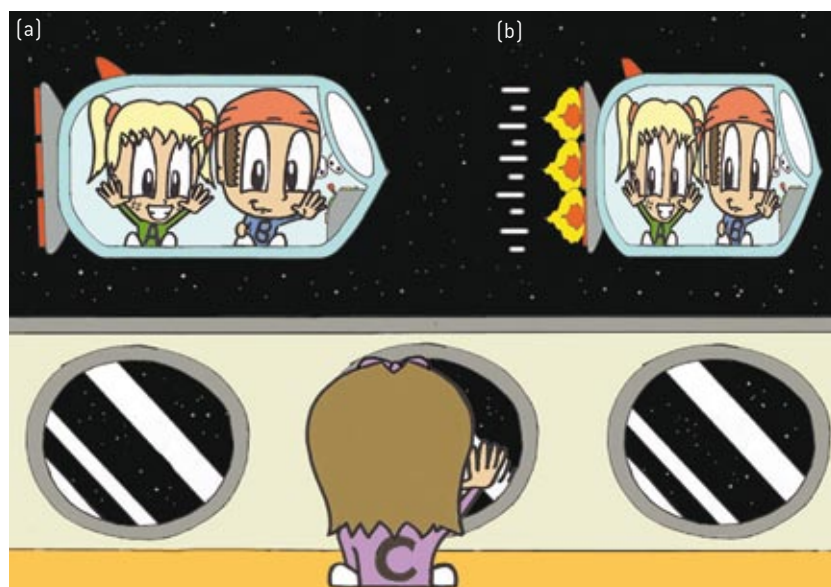


Figure 6.28 (a) Chloe watches as Anna and Ben prepare for blast off. (b) As Anna and Ben speed past, Chloe sees their rocket ship foreshortened, but the other dimensions remain the same.

The proper length is the *distance between two points whose positions are measured by an observer at rest with respect to the two points*. Again, as Anna and Ben can read their measuring tape at either end of the carriage and are at rest with respect to the train, their measurement is the proper length while Chloe's measurement will be of the contracted length. Remember that length contraction occurs only in the direction of travel, not in any perpendicular direction. To Chloe, Anna and Ben's spaceship will appear foreshortened, but its width and height will remain unaltered.

These two simple expressions—the length contraction equation and the time dilation equation—are at the core of the special theory of relativity. Together they tell us about spacetime, the four-dimensional world that we inhabit.

Worked example 6.4A

In Worked example 6.3A we were left with the question as to how the space traveller to Vega (at 99.5% of c) apparently covered a distance of 25 light-years in just over 5 years. Didn't this mean that he was travelling faster than light?

Solution

The key is that the traveller sees the galaxy going by in the opposite direction at 99.5% of c ($\gamma = 10$ at this speed) and so it appears contracted. From his frame, then, the distance he covered was not 25 light-years, but only $25/\gamma = 2.5$ light-years. Remember that the Earth-bound observers saw the trip take just over 25 years although they 'saw' that his clock took just 2.5 years. The problem in the question is that we are asking about a distance measured in the Earth's frame, but a time measured in the traveller's frame. A speed can only be given when the two are measured in the same frame. Neither observer saw a trip faster than light in their own frame!

Four-dimensional spacetime

Earlier we discussed the idea that our world might actually be four-dimensional rather than having only the three dimensions we normally perceive. But what does this really mean? When we give the position of something we *must* give three coordinates. Remember that we often assume one of them. If we say that our house is at D8 on map 26 of the street directory, the D and the 8 represent two of the coordinates. (The map number and directory only tell us about the system of coordinates we are using.) The third coordinate in this system is assumed to be ground level. If we live in a tall apartment building we would need to give the floor level as well. These three coordinates are interrelated. The D by itself cannot specify our location without the 8 and the map details.

So what would it mean to live in a four-dimensional world? In one sense we already do. If we specified the location of our house on a party invitation we would also need to give the time. However, the time is really quite independent of the location. We could, for example, decide to move the party to a friend's house and simply tell our friends to go there instead. Or we could decide to have it our house but on a different day. Changing either the time or the location does not affect the other, they are *independent* of each other. Compare this situation with giving the details of a new location. It would be no good just giving one of the map coordinates by itself, we need to give both—the map coordinates are *interdependent*. The time of our party is not affected at all by what street directory we use, but imagine what would

Physics file

The Lorentz contraction factor becomes so close to one for values of v less than about $0.0001c$, that the expression

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

can't be used with a normal calculator. Fortunately, there is a simple way to find the value of γ for speeds less than about 1% of c . A binomial expansion of the term $(1 - x)^n$ tells us that, provided $x \ll 1$, $(1 - x)^n = (1 - nx)$. For the Lorentz factor, $x = (v/c)^2$ and $n = -\frac{1}{2}$, and so $\gamma = 1 + \frac{1}{2}(v/c)^2$. Thus the part of the factor greater than one can simply be found from $\frac{1}{2}(v/c)^2$.

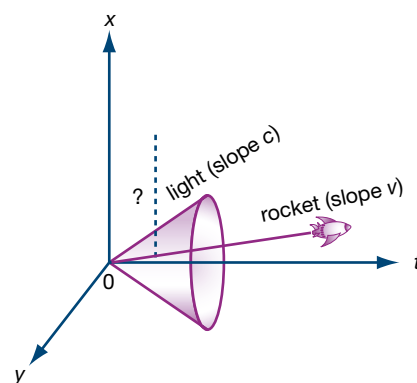


Figure 6.29 It is difficult to represent four different dimensions on two-dimensional paper. Here we have simply left out one of the space dimensions in order to include the time dimension. The cone represents a flash of light that occurred at time and space zero. It spreads into spacetime at speed c . The line represents the trip of a fast rocket ship coming from the same place and time. Can you see why the dotted line represents an impossible trip?

happen if we gave the letter coordinate of our house (D) from the Melway street directory but the number coordinate (8) from the UBD directory! These coordinates are actually interdependent.

Einstein realised that, contrary to what we normally experience when sending or receiving party invitations, location coordinates and time co-ordinates are *not* independent; they are in fact *interdependent*. He therefore suggested that we think of spacetime, a world of four dimensions: three space dimensions and one time dimension. In spacetime all four dimensions are interdependent. As we have seen, motion through space comes at the expense of motion through time.

Let's rejoin our space travellers for a party at Anna's space station, X-ray. Ben, who lives on planet Yerkes, knows that, relative to him, space station X-ray is in a yerkescentric orbit and thus at a fixed distance from him. On the other hand, Chloe, who lives on space station Zulu, is across the galaxy from the others and travelling at $0.1c$ (10% of the speed of light) relative to X-ray. The problem, of course, is how to have everyone arrive at the same time. Ben should not have too much trouble—after all, as he watches Anna's clock (or picks up its radio signal), it at least ticks at the same rate as his own. If he synchronises his clock with Anna's, it will keep the same time.

But Chloe has a real problem. As she watches Anna's clock, it seems to be going slow, relative to hers. She could synchronise hers one day, but in a week it will be well ahead of Anna's again. So when in fact is the party? And how long is going to elapse on X-ray while she travels from space station Zulu?

We will leave our travellers to sort out their own dilemmas—it is all getting too complicated! The point is that time and space become mixed up together when we start thinking in terms of vast distances and super high speeds. We can no longer issue party invitations with a simple place and time to all our friends spread around the galaxy. It is not just a matter of synchronising watches, or logging on to the intergal-net to check universal time. *There is no universal time*. To repeat: Time goes at different rates for people travelling at different relative speeds. Space and time coordinates are interdependent.

Physics in action

The invariant spacetime interval

You may be wondering whether there is anything that all observers *would* agree on, or to put it another way; have Anna, Ben and Chloe got any hope of getting together for that party? The answer is yes! It is called the *invariant spacetime interval*. This material is beyond what is required for our course, but is provided for your interest.

Consider a simple analogy. While the position of an object depends on our system of coordinates, the distance between two objects does not. We can walk from a house 3 km down the road to another 5 km down the road. These positions depend on the starting point (the origin of our coordinates), but the distance—the 3 km walk—does not. The distance is said to be *invariant* under different coordinate systems. We have already seen that in spacetime, distance and time are *not* invariant, they both depend on the relative speed of the observer. In a set of *xyz* coordinates the distance between

two points is given by the three-dimensional version of Pythagoras's theorem:

$$d^2 = x^2 + y^2 + z^2$$

We choose one point to be the origin, for simplicity. In relativity, we often speak of 'events'. Events are specified by a position and a time (like Anna's party). Two events may occur in the same place at different times, the same time but in different places, or different times and different places. In normal situations we can speak of the distance between events (given by the equation above) and the time between events (simply $t_2 - t_1$). These are quite independent of each other.

As we now know that space and time are connected, we can expect that things will not be quite that simple in spacetime! However, it turns out that in spacetime the

quantity $x^2 + y^2 + z^2 - c^2t^2$ plays a somewhat similar role to d^2 in normal space. As ct is the distance light travels in time t , this quantity could be seen, more or less, as a four-dimensional version of Pythagoras's theorem. We call the square root of this quantity (the analogy of d in Pythagoras's theorem) an *invariant spacetime interval (isti)* between two events. All observers in inertial frames will agree that this quantity is the same, no matter what their relative velocities. We can see that we have a kind of trade-off between space and time here: if the distance part of the expression is greater, the time will need to be longer to preserve the value of the *isti*. Let's consider an example.

Ben pops the cork at Anna's party and it falls back on top of the bottle. Anna and Ben see these two events occur at the same place, but separated by a short time interval t_0 (this is the proper time). Thus the quantity $x^2 + y^2 + z^2$ is zero, but

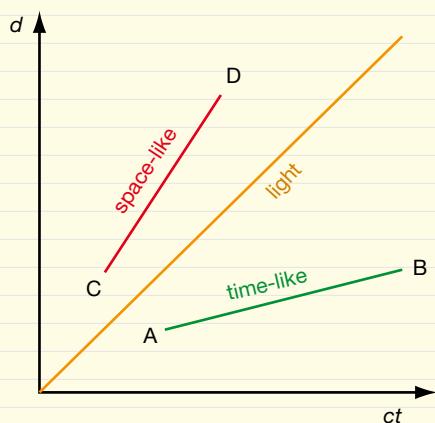


Figure 6.30 Events can have space-like or time-like separation. On this graph, space has been simplified to one dimension and the other axis is ct , the distance light travels in time t , thus the slope of the light line is 1. Events separated by a line with a slope less than 1 (A–B) have time-like separation and those with a slope greater than one (C–D) have space-like separation.

there is a small t_0 and so the *isti* has the value $-c^2t_0^2$. Chloe, who was speeding by the party, saw the cork fall at a different position—still on the bottle, but that was moving as well in her frame. For Chloe then, $x^2 + y^2 + z^2$ has a positive value. To preserve the negative value of the *isti* this means that the time t (for her) will need to be greater. This agrees with the fact that we already know that her time will be related to Anna's time by $t = \gamma t_0$ and so will be greater by the factor γ .

Before leaving the *isti* it is worth noting that its value can be positive, negative or zero. The cork popping at Anna's party involved two events which occurred at the same place (as seen by the party-goers), but at different times. The value of the *isti* was therefore negative as the distance part of the expression was zero. For their next trick our three space travellers are at three different locations, but all at rest with respect to each other. Anna is halfway between Ben and Chloe and sends them a simultaneous signal which causes them both to jump for joy. Because she is halfway between, we don't need to worry about any look-back time effects. Anna sees them both jump at the same time. These two events are therefore separated in space, but not in time (so the c^2t^2 term is zero). Thus our *isti* is positive this time.

Comparing these two different pairs of events, the cork popping was of the type where a single observer could be present at both (the pop and the falling back). On the other hand, no matter how fast we travelled we could not be present at both Ben's and Chloe's jumps. We can generalise these situations. When our *isti* is negative we can be present at both events, they are separated by a time that allows us to travel between them. When the *isti* is positive, however, we cannot travel fast enough to be present at both. And what if *isti* is zero? You have probably guessed already that this is the situation where only light can make it between the two events. (If $d^2 - c^2t^2 = 0$, then $d = ct$.) In relativistic terms the cork popping had a *time-like separation* and the jumps had a *space-like separation*.

Arguments such as these led Einstein to one of his most famous conclusions. Travel at greater than the speed of light is not possible. In fact, not only travel, but the transmission of any sort of information is not possible at speeds greater than c .



6.4 summary

Time and space

- 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, by the factor γ . Einstein's length contraction equation is given by:

$$l = \frac{l_0}{\gamma}$$

- The *proper time*, t_0 , is the time measured by an observer in the same frame of reference as two events.
- The *proper length*, l_0 , is the length measured by an observer at rest with respect to two points.
- Einstein said that we live in a four-dimensional world of spacetime in which space and time are interdependent. Motion through space comes at the expense of motion through time.



6.4 questions

Time and space

- 1 Our earlier simple determination of the apparent difference between Anna's and Chloe's measurements of the time for the flash of light to return to the centre of the carriage was incorrect because:
 - A we made a mistake in the algebra
 - B we didn't allow for the look-back time involved
 - C we didn't allow for the reduced speed of light in the moving frame of reference
 - D we didn't realise that the length of the carriage would be reduced because of its motion in Chloe's frame.
- 2 If we observe a speeding rocket ship we find that (one or more):
 - A its clocks seem to be going fast
 - B its clocks seem to be going slow
 - C its length appears to be shorter than normal
 - D its length appears to be longer than normal.
- 3 If we observe a speeding rocket ship we find that (one or more):
 - A all its dimensions appear smaller
 - B all its dimensions appear normal
 - C its width appears smaller
 - D its width appears normal.
- 4 As Anna and Ben speed along at $0.9c$ (relative to Chloe) in their rocket ship, Anna is holding a metre rule parallel to the direction of their velocity.
 - a What is the length of the rule as measured by Ben?
 - b What is the length of the rule as measured by Chloe?
- 5 Now Anna turns the ruler through a right angle so it points across the ship.
 - a What is the length of the rule as measured by Ben?
 - b What is the length of the rule as measured by Chloe?
- 6
 - a At what speed would the rocket ship be going if Chloe observed it to be half its normal length?
 - b Chloe then observed the rocket ship to accelerate so that its length halved again. Did that mean that it doubled its speed? To what speed did it accelerate?
- 7 In which of these cases will the measurement of length be the proper length?
 - A Ben measures the length of his speeding rocket ship with a tape measure.
 - B Chloe measures the length of Ben's speeding rocket ship from her space station.
 - C Ben measures the length of Anna's speeding rocket ship from his own mini-rocket travelling at the same velocity as Anna's.
 - D Chloe takes a photo of Anna's ship as it passes a fixed length scale and finds the length from the photo.
- 8 Which one or more of the following conditions is sufficient to ensure that we will measure the proper time between two events? We must:
 - A be in the same frame of reference
 - B be in a frame of reference which is travelling at the same velocity
 - C be stationary
 - D not be accelerating with respect to the frame of the two events.
- 9 Why is travel on a vertical line on the spacetime graph shown in Figure 6.29 an impossibility?
- 10 When calculating the value of γ for speeds quite a bit less than that of light, it becomes difficult for a normal calculator to handle the large number of digits involved. However, as shown in the Physics file on page 217, the binomial theorem enables us to approximate the value of γ with the expression: $\gamma \approx 1 + \frac{1}{2}(v/c)^2$. Compare the values obtained using this expression with the correct values for velocities of 1%, 10%, 25% and 50% of the speed of light. At what point is the expression no longer satisfactory?
- 11 The space shuttle orbits the Earth at around 8000 m s^{-1} . Try to find the value of gamma (γ) for the space shuttle using the correct expression, and then by using the binomial expression in Question 10.

6.5 Momentum, energy and $E = mc^2$

The Lorentz factor, γ , rapidly increases as the speed, v , gets closer to the speed of light, c . At 99.9% of the speed of light, γ has a value of about 22 and so anything moving at that speed (relative to us) will appear to have shrunk to 1/22nd of its normal length. As we watch the action inside a spaceship travelling at that speed, it would appear to be going very slowly, only 1 minute for every 22 of ours. The closer the speed of our spaceship gets to the speed of light, the more the Lorentz factor skyrockets towards infinity. So what happens *at* the speed of light? According to Einstein's equations the length of the spaceship shrinks to zero and time inside it appears to stop altogether. Einstein took this to mean that, in fact, it is *not possible* to reach the speed of light in any real spaceship. However, the difficulties with time and length for our spaceship were not the only reasons Einstein came to this conclusion (for example, see the Physics in action on page 219).

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. Although his analysis is beyond our scope, we can get a feel for his approach if we take some short cuts.

The acceleration of any object depends (inversely) on its inertial mass, the mass that appears in Newton's famous second law, $F = ma$. Newton originally stated this law in a form more like what we know as the impulse equals change of momentum principle: $Ft = \Delta mv$. Now we see that time is involved, but at relativistic speeds we know that time becomes rather rubbery!

Imagine a rocket ship accelerating from rest in our frame. We can say that the momentum of the ship will be given by $mv = Ft_0$ where t_0 is the time in the ship's frame. In our frame the time is $t = t_0\gamma$ and so $t_0 = t/\gamma$. A little rearranging leads to $Ft = \gamma mv$. That is, the impulse as we see it in our frame is equal to γ times the Newtonian momentum. We can interpret this as meaning that the momentum can be seen as: $p = \gamma mv$, or $p = \gamma p_0$ where p_0 is the momentum (mv) as we would see it in classical mechanics. While we have taken some short cuts to reach this result, the result itself is perfectly valid. In 1909, 4 years after the original paper, Einstein produced the *relativistic momentum equation*.



The **RELATIVISTIC MOMENTUM EQUATION** is given by:

$$p = \gamma p_0 \text{ or } p = \gamma mv$$

We can now see why our rocket ship cannot reach the speed of light. 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! Worked example 6.5A illustrates the point.

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.

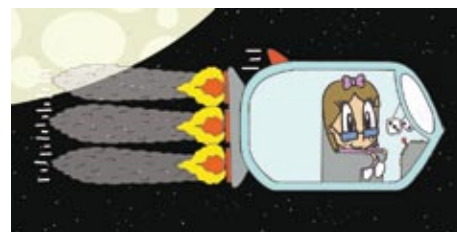


Figure 6.31 This rocket ship is doing $0.99c$ and accelerating, so why can't it accelerate to $1.00c$?

Worked example 6.5A

What impulse is needed to bring a rocket ship of mass 1000 kg to 0.99c? If this impulse is applied again, what speed will the ship reach?

Solution

The momentum of the ship at 0.99c, and hence the impulse required, is given by $p = \gamma mv$.

At 0.99c, γ is 7.09 and so:

$$p = 7.09 \times 1000 \times 0.99 \times 3 \times 10^8 = 2.1 \times 10^{12} \text{ kg m s}^{-1}$$

If this impulse is applied again, the total momentum will double and hence become:

$$4.2 \times 10^{12} \text{ kg m s}^{-1}$$

As v appears in γ as well as mv , it is difficult to solve the equation for v . We can, however, make a simplifying assumption that v will be very close to c and so find γ . Thus:

$$\gamma = \frac{p}{mv} = \frac{4.2 \times 10^{12}}{1000 \times 3 \times 10^8} = 14$$

As can be seen from the graph in Figure 6.25 (or checked by calculation) this value of γ implies a speed of almost 0.998c.

One interpretation of this strange behaviour of momentum is that it appears that the inertial mass of the rocket ship is increasing with the Lorentz factor. In other words, we can interpret the γm part of the relativistic momentum equation to be the *relativistic mass*.



RELATIVISTIC MASS is given by:

$$m = \gamma m_0$$

where m_0 is the rest mass of the rocket ship [the mass measured in its own frame of reference].

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.

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:

$$E_k = \frac{1}{2}mv^2 = \frac{1}{2}pv$$

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:

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

Although it is not very obvious from this expression, if the velocity (which is hidden in the γ term) is small, this expression actually reduces to our familiar $E_k = \frac{1}{2}mv^2$. 'Small' in this context means small compared with c .

Physics file

As we watch a rocket ship travelling at 0.99c (speed U) it fires a small ship at 0.02c relative to it (speed v). Isn't the small ship moving at 1.01c? No! First we need to be careful to specify in which frame of reference we are measuring the speeds. The rocket ship has speed U in our frame while the small ship has speed v in the frame of the rocket ship. (Capital letters for our frame, small for the rocket frame.) Because of length contraction we see the small ship fired at much less than 0.02c. Einstein showed that in these cases the speed (V) of the small ship in our frame is given by:

$$V = \frac{U + v}{1 + \frac{Uv}{c^2}}$$

You can use this expression to show that we see the small ship travelling at 0.9904c.

Physics file

Mathematically curious students will not find it too difficult to show that Einstein's expression $E_k = (\gamma - 1)m_0 c^2$ does indeed reduce to $\frac{1}{2}mv^2$ when v is considerably smaller than c . What is needed is the binomial expansion:

$$(1 - x)^n = (1 - nx) \text{ if } x \text{ is } \ll 1$$

If this is applied to $\gamma = (1 - v^2/c^2)^{-\frac{1}{2}}$, you find the result you are looking for.

But even for speeds up to 10% of c , the normal expression is accurate to better than 1%.

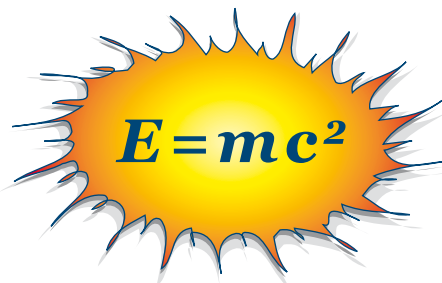
Einstein's expression can be rewritten as $E_k = \gamma m_o c^2 - m_o c^2$, which in turn can be rearranged as $\gamma m_o c^2 = E_k + m_o c^2$. Einstein interpreted this expression as being an expression for the *total energy*, $\gamma m_o c^2$, of the object. The right-hand side appeared to imply that 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. Now as γm_o is the relativistic mass, m , the total energy can simply be written as $E_{\text{tot}} = mc^2$. You will have seen this equation before!



TOTAL ENERGY is given by:

$$E_{\text{tot}} = E_k + m_o c^2 \text{ or simply } E = mc^2$$

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 (9×10^{16} in SI units) 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.

Worked example 6.5B

How much heavier is a 200 g cup of boiling water than the same cup of iced water? [Given that the increase in heat required is $mc\Delta T$, where m is the mass, c is the heat capacity $4.2 \text{ J g}^{-1} \text{ }^\circ\text{C}^{-1}$, and ΔT the temperature rise.]

Solution

To heat the 200 g of water by 100°C requires $200 \times 4.2 \times 100 = 84\,000 \text{ J}$. The mass equivalent of this energy is given by:

$$m = \frac{E}{c^2} = \frac{84\,000}{9 \times 10^{16}} \approx 10^{-12} \text{ kg}$$

So don't bother trying to weigh hot and cold cups of water!

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



Figure 6.32 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.

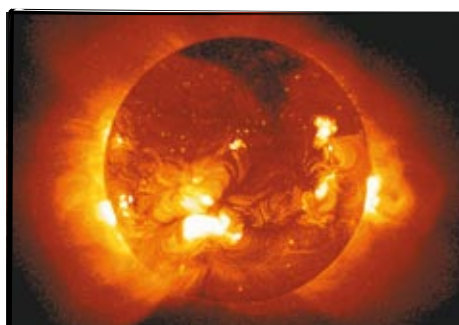


Figure 6.33 Nuclear fusion in the Sun results in about 4 million tonnes of mass being lost with the energy radiated from the Sun every second. This photo is taken in soft X-ray light.

Physics file

The Sun is losing weight! The Sun's energy is produced by the nuclear fusion of hydrogen atoms forming helium atoms. Every second, about 600 million tonnes of hydrogen atoms is converted to helium atoms, which have a total mass about 4 million tonnes less than that. Each second, this 4 million tonnes of mass has gone off as mass-energy radiated into space. Don't worry about the Sun disappearing, however. Its mass is around 2×10^{30} kilograms, so at this rate it would take hundreds of billions of years to use it all up. However, other processes, to do with new nuclear reactions forming, are likely to result in the Sun exploding to a red giant in a mere 5 billion years!

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.

Was Einstein right?

It is said that when Einstein heard of one of the first pieces of experimental evidence that supported his theory he did not seem all that elated. When someone later asked him why he was not as excited as many of the others in the room were, he simply said that if it had not turned out to be true 'that would be too bad for God'. This was neither blasphemy nor arrogance. He had a deep conviction that the beauty and simplicity of relativity meant that it had to be right. For most people, however, experimental proof was still important!

By the end of the 20th century the experimental proof was overwhelming. Relativity is one of the most successful theories in physics. Its predictions have been shown to be true over and over again. No twins have been on long space journeys, but many accurate clocks have been flown around the world in aeroplanes and in satellites. In every case, the travelling clock returns to its stay-at-home 'twin' having lost just the amount of time Einstein's equations predicted. The whole global positioning system (GPS) is based on relativity theory. If Einstein was wrong, the GPS system would be out by miles—literally.

The increase in mass of a space shuttle travelling at 20000 m s^{-1} is both minute and impossible to measure. However, for many decades now, particles such as electrons, protons and ions have been accelerated to speeds very close to that of light. It is quite easy to measure the mass of such a particle: simply send it through a magnetic field, which will deflect its path. The amount of deflection depends on various quantities such as the magnetic field, but also on the mass of the particle, a heavy particle deflecting less than a light one. Needless to say, the mass increase with speed is just as relativity predicts. If it were not for the relativistic mass increase of the electrons speeding around at almost the speed of light in the new Australian synchrotron, the machine could be just a few *centimetres* across instead of 70 m in diameter!

We don't need to go to such exotic machines for relativistic effects, however. The electrons in a cathode ray TV set are actually travelling at around 30% of the speed of light and the magnetic deflection that paints the picture on the screen would give us a very fuzzy and small picture if the relativistic effects on the deflection were not taken into account.

One interesting proof of relativity came in the observation of subatomic particles called *muons* created by collisions of cosmic rays with atoms high in the Earth's atmosphere. Muons have a very short life. On average, half of them only live for $2.5 \mu\text{s}$. It is easy enough to measure their speed as they plunge down through the atmosphere—it is about $0.99c$, almost the speed of light. Now at that speed, a muon that lives $2.5 \mu\text{s}$ will travel $3 \times 10^8 \times 2.5 \times 10^{-6} = 750 \text{ m}$, not nearly far enough to reach the surface of the Earth, typically 15 km below. At that speed it would take close to $50 \mu\text{s}$ to reach the ground. However, over 10% of the muons created at that height *do reach* the ground. The explanation, as you have probably guessed, is that

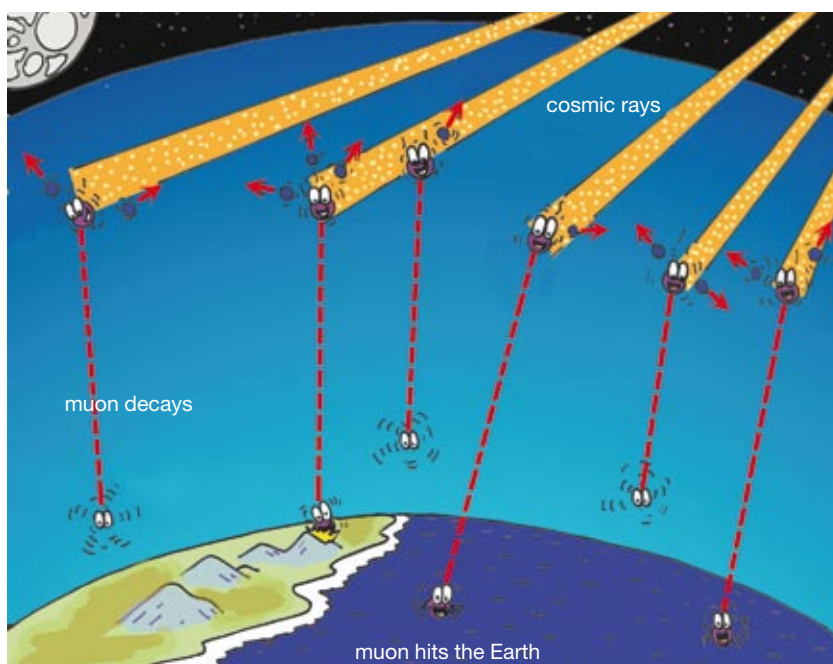


Figure 6.34 Muons should not reach the Earth's surface according to Newton; however, many do.

time is passing much more slowly for the muons—as we see it. At $0.99c$, the Lorentz factor is 7.1 and so the muon half-life (in our scheme of things) is nearly $18 \mu\text{s}$. This is still not long enough for the average muon to reach ground level, but a significant number of muons (12.5%) will live for three half-lives, which is long enough to reach the ground.

Einstein's curiosity was aroused by a very simple phenomenon—magnetism. His special theory of relativity arose largely from his attempts to explain an apparent problem with magnetic forces. Conventional theory said that the magnetic force depended on the speed of moving charged particles. But that meant that the force should disappear in the frame of reference in which the particles were at rest! Clearly it did not, but classical physics had no explanation. Relativity does explain magnetism and anyone who has played with magnets, or used an electric motor, has seen a direct confirmation of relativity theory! (See the Physics in action, page 227, for more details.)

Every now and then one hears of an experimental observation that seems to have cast doubt on the theory of relativity—something that seems to travel faster than light, or the speed of light is slowing down, for example. Despite all these, relativity is one of the most well-tested and accurate theories ever put forward in physics. Certainly no physicist will claim that it is the final word, that no modification will ever be discovered, but the attention any apparently inconsistent observation receives is an indication of the strength and importance of the theory. When a modification is discovered it will be as revolutionary as the theory Einstein put forward in 1905.

Simplicity and beauty

Einstein based his ideas on a conviction that two basic principles of physics *had* to be right. The first was the principle of relativity as put forward by Galileo and Newton. The second was Maxwell's theory of electromagnetism, which in turn was based on the work of Michael Faraday. All of these great figures in physics had a belief in the beauty and elegance of the laws of



Figure 6.35 The relativistic increase in mass of the electrons in a TV tube needs to be taken into account by the engineers who design TVs or the picture would be small and distorted.

nature. It seemed to them that the world is designed on simple but profound principles which make sense to us. Einstein put it this way: 'The most incomprehensible thing about the universe is its comprehensibility.'

In his book *The Elegant Universe* physicist Brian Greene tells us:

Einstein was not motivated by the things we often associate with scientific undertakings, such as trying to explain this or that piece of experimental data. Instead, he was driven by a passionate belief that the deepest understanding of the universe would reveal its truest wonder: the simplicity and power of the principles on which it is based. Einstein wanted to illuminate the workings of the universe with a clarity never before achieved, allowing us all to stand in awe of its sheer beauty and elegance.

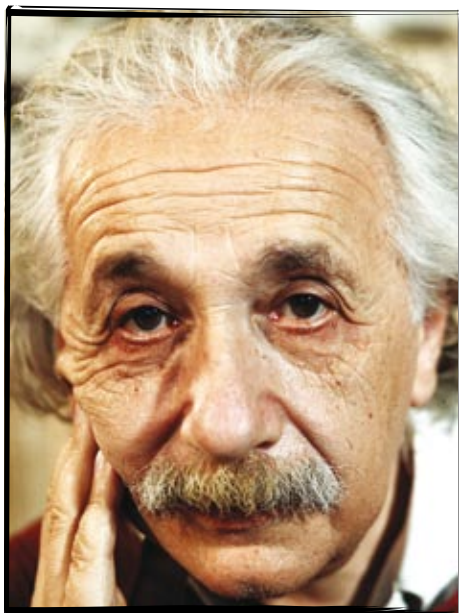


Figure 6.36 Albert Einstein (1879–1955).

To most of us, Einstein's relativity seems complex and difficult at first. But to those who make the effort, after a while it becomes apparent that it has an elegance, indeed a simplicity, which convinces one that it just has to be true. This is not a matter of blind faith; the physicist always looks to experiment as the ultimate justification of a theory. Indeed, the conviction that good theory has an element of beauty about it is based on experience. Newton's laws of motion describe a huge array of situations in just three simple statements. In a real sense they have a beauty about them just as undeniable as that of a Renoir painting. The same can be said for Maxwell's electromagnetic equations. Furthermore, good theories are always open to modification, just as Newton's mechanics was modified by Einstein's relativity. The future may, indeed probably will, see a grander, all-encompassing theory of space, time and matter, but we can be sure that Einstein's relativity will be a part of it. Certainly it will be seen as an important stepping stone along the way.

Consider one example of the simplicity to be discovered in relativity. The speed of light is constant because the two constants that describe the way electricity and magnetism behave are constant in any inertial frame of reference. These constants also determine many of the properties of electric circuits. If these quantities changed with our velocity (as, for example, the



Figure 6.37 If the speed of light was not constant, modern electronic devices would not work when in motion.

speed of sound does) then all our computers, radios, telephones and other modern paraphernalia would only work at one particular velocity; and don't forget the Earth is continually changing its velocity. Imagine the chaos! The constancy of the speed of light may be a little difficult for us to comprehend, but it is a simplifying principle.

We have also seen that an array of different and apparently unrelated quantities—space, time, mass and energy—is reduced to two quantities: spacetime and energy. These sorts of simplifications occur right through relativity. For example, in general relativity even acceleration and gravitation are combined. Certainly it is hard for those of us just beginning with relativity to fully appreciate the elegance and beauty of the theory. Hopefully, the little glimpses we have seen in the course of our rather sketchy outline of the subject will help us to appreciate one of the most profound achievements of the human mind.

The most beautiful thing we can experience is the mysterious. It is the source of all true art and science.

Albert Einstein

Physics in action

Magnetism and relativity

Magnetism is a relativistic phenomenon! We sometimes think that relativity only applies to ultrafast rocket ships or exotic particles in accelerators, but, in fact, before relativity there was no sound explanation of the magnetic forces between electric currents moving at speeds of only millimetres per second. The 19th-century physicists knew that there was a problem with the theory of magnetism. It was well known that a moving charge (and hence a current) in a magnetic field experiences a force that is directly proportional to the velocity of the charge ($F = qvB$, see Chapter 9, page 347). Indeed, this is the force that drives all the electric motors of the world. But how could we have a velocity-dependent force without contravening the Galilean principle of relativity?

The problem was that in the frame of reference of the moving charge, the velocity of the charge, and hence the magnetic force on it, should be zero. This was clearly not the case, however. So was there something special about the frame of reference after all, despite the Galilean principle of relativity? At the beginning of the 20th century this was one of the unsolved mysteries of physics.

Consider two electric currents moving in the same direction in two similar parallel wires. We know that in this situation there is a magnetic force of attraction between the two wires; a simple experiment can confirm this (see section 9.2). Now, imagine a moving electron in one of the wires. It 'sees' the magnetic field created by the current in the other wire and 'feels' a magnetic force towards it. But what if we observe this situation from a frame of reference moving at the same velocity as the electrons, which is literally only a snail's pace? In this case, the electrons are all at rest and so there should be no force!

Now, the force between two objects cannot depend on the frame of reference. Either they will get closer or they won't, and that doesn't depend on how we look at them. We know that wires with parallel currents *do* get closer, so we know there is a fault in our physics if it says there is no



Figure 6.38 These three diagrams represent two similar electric currents. (a) The upper wire has moving electrons and fixed positive charges. The lower current is represented by a stream of electrons as the positive charges play no role in our story. This is the conventional view; the stream of electrons and the current are attracted because the moving electrons in each create a magnetic field and experience a force from the field created by the other. (b) From the frame of reference of the electrons, which are now at rest, conventional theory says that although the positive charges are moving, the electrons experience no force. This cannot be correct! (c) This is the relativistic version. Here, from the electron's frame, the positive charges are moving and their space is contracted, making them appear to be closer together. There is thus a net attraction between the positive charges and the electron stream. The contraction is wildly exaggerated here, but the number of particles in a real current is also much greater!

force between them. This type of situation is not uncommon in science. In fact, it is one of the ways in which science progresses. Einstein was very aware of the problem of electromagnetism and indeed his famous 1905 paper starts with a discussion of just this problem.

Normally, when discussing electric currents we simply think of the moving electrons and ignore the huge numbers of positive and negative charges that are at rest in the wire. The very good reason for this assumption is that the total negative charge of the conduction electrons is almost equal in magnitude to the total positive charge of the positive ions (the atoms with the remaining electrons); so the wire as a whole is neutral. (Any small overall charge because of a positive or negative voltage on the wire is negligible.) As in any problem, however, we need to look at our assumptions very carefully.

Our hypothetical moving electron actually 'sees' a huge electrostatic (Coulomb) force towards all the positive charges in the other wire, but this is balanced by the equally huge repulsive force from all the negative charges in that wire—or is it? Classical theory certainly says that these forces should balance. However, relativity tells us to be careful where we have relative motion. The moving electron actually sees all the moving electrons in the other wire at rest, relative to itself, but it sees all the positive charges moving in the opposite direction. Now that means that the positive charges will appear—to our moving electron—contracted or, more particularly, the space that they occupy will appear to be shortened in the direction of their motion. And so their density—the number of positive charges in a metre of wire—appears greater than the density of the negative charges. There is, therefore, an imbalance in the Coulomb force between our electron and the negative and positive charges in the other wire. The electron sees more positive charges than negative charges and so is attracted to the other wire!

The obvious comment to be made here, however, is that the speed involved—a few millimetres per second—is so small that the Lorentz contraction should be totally negligible. However, it is important to remember that we are dealing with *huge* numbers of charges. It is worth doing a simple calculation as an illustration of the forces we are dealing with:

If we could take all the conduction electrons out of a piece of wire and separate them by 1 cm from all the positive ions remaining, what would be the force between them? A piece of

copper wire 10 cm long and with a cross-section area of 1 mm² has a mass of about 1 g. As its atomic mass is 64 atomic mass units, there are about 10²³ atoms. Let's assume that one electron from each atom is taken out and placed 1 cm away from the remaining positive ions. The total charge of the electrons is then 10²³ × 1.6 × 10⁻¹⁹ ≈ -2000 C, and there will be a charge of +2000 C on the positive ions. The force between the electrons and the rest of the wire is approximately given by the Coulomb force:

$$F = \frac{kQq}{R^2} = \frac{9 \times 10^9 \times 2000^2}{(10^{-2})^2} \\ \approx 4 \times 10^{20} \text{ N}$$

This force is about the same as the gravitational force that holds the Moon to the Earth! Or put another way, it is the weight of over one hundred billion supertankers! The electrical force between the particles in a piece of wire is absolutely huge!

Perhaps the point of this calculation is becoming clear? The electrical force on the moving electron in our wire in the magnetic field of another wire is a very delicate balance between two enormous forces—that from the positive protons in the wire and that from the negative electrons in the wire. Clearly any slight imbalance in those two forces will have an enormous effect. (You might like to calculate the force on just one electron from the positive charges in the wire in the previous example.) Although the Lorentz contraction is very slight, it is enough to produce a very small imbalance in the force on our electron. You can confirm that the Lorentz factor differs from 1 only in about the 23rd decimal place, but if that figure is multiplied by something like 10²⁵ N, we end up with a normal sort of force; in fact, we end up with what we call the *magnetic force*.

So the magnetic force is actually a normal Coulomb force that results from a slight imbalance in the huge forces between all the protons and all the electrons in a wire in which many of the electrons are moving. Again we find that relativity is actually a simplifying principle. What were thought to be two different, but related, forces are actually different aspects of the one *electromagnetic force*. The fact that all the electric motors we use every day work so easily and efficiently, whether we are at home listening to a CD or flying around the world in an aircraft, is excellent evidence of the validity and relevance of Einstein's great theory.



6.5 summary

Momentum, energy and $E = mc^2$

- Because the Lorentz factor, γ , approaches infinity at the speed of light, the length of a moving object approaches zero and time comes to a standstill as its speed approaches c .
- Relativistic momentum also includes γ and hence as more impulse is added, the mass seems to increase towards infinity as the speed gets closer, but never equal, to c .
- Einstein found that the total energy of an object was given by $E_{\text{tot}} = E_k + E_{\text{rest}} = mc^2$ where $m = m_0\gamma$. The kinetic energy is given by $E_k = (\gamma - 1)m_0c^2$.
- The rest energy is energy associated with the rest mass of the object $E_{\text{rest}} = m_0c^2$ 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.
- A vast amount of experimental evidence now supports Einstein's special theory of relativity.



6.5 questions

Momentum, energy and $E = mc^2$

The following information applies to questions 1–3.

We are watching a rocket ship that is approaching the speed of light in our frame of reference.

- 1 Describe what we see as the speed of the rocket ship increases.
- 2 Inside the rocket ship the crew will notice one or more of the following:
 - A their clocks slowing down
 - B their ship getting shorter
 - C the Earth becoming foreshortened
 - D nothing unusual in their ship.
- 3 While the mother ship is travelling at $0.9c$, a small explorer ship leaves the mother ship and accelerates up to $0.2c$ relative to the mother ship. Which of the following suggest why we do not see the explorer ship travelling at $1.1c$?
 - A Because of length contraction, the distance the small ship travels appears less to those of us on board than to those on the mother ship.
 - B Because of time dilation, the time it takes to move each metre is longer to those of us on board than to those on the mother ship.
 - C The explorer ship cannot actually achieve a speed of $0.2c$ relative to the mother ship because the mother ship is already travelling at $0.9c$.
 - D The small ship does actually travel at $1.1c$, but it is too fast for us to see.
- 4 A rocket ship travelling at $0.99c$ turns on its motor and accelerates.
 - a Will we see it accelerate?
 - b Will the crew experience acceleration?
- 5 Why is it that while the crew of a rocket ship that is accelerating can experience considerable acceleration forces, we may see the rocket ship accelerate only very slightly?
- 6 Briefly explain why, as we watch a rocket ship accelerating under a constant force from its motor, it can never reach the speed of light.
- 7 Why is it that while thousands of tonnes of coal must be burnt each day to produce enough power for a city, only a few grams of nuclear fuel is lost in a nuclear power station with the same output?
 - A The nuclear fuel converts mass into energy but the coal does not.
 - B The electrons in uranium atoms are held more tightly than those in carbon atoms.
 - C Nuclear power stations burn their fuel more efficiently than coal stations.
 - D The forces in the nucleus are very much stronger than those between atoms.
- 8 Melbourne uses around 200 000 GJ of energy each day. This is provided (mostly) by burning about 100 000 tonnes of coal each day.
 - a What loss would there be in the mass of the reactant products of the burnt coal?
 - b What loss would there be in the mass of the uranium fuel if this amount of energy was produced by a nuclear reactor?
- 9 A kilogram of coal will produce about 25 MJ of heat energy. Very approximately, how long would the Sun last if it was burning coal instead of nuclear processes? (See Physics file on page 224.)
- 10 Discuss the statement: Relativity is only a theory and may be overturned tomorrow.

chapter review

Multiple-choice questions

- Newton's second law of motion states that the acceleration of a body is proportional to the force applied. In which way does this law strongly support the principle of relativity?
 - Because a force produces an acceleration, it results in a certain velocity. This velocity is quite independent of the frame of reference.
 - Because a certain force produces a certain velocity which is independent of the frame of reference.
 - Because a force produces an acceleration, it causes a change in velocity. The actual change of velocity is dependent on the frame of reference.
 - Because a force produces an acceleration, it causes a change in velocity. The actual velocity does not matter and so the law is quite independent of the frame of reference.
- Galileo tried to measure the speed of light by having an assistant uncover a lamp on a hill about 2 km away when he saw the light from Galileo's lamp. Which of the following is a reasonable estimate of the minimum value for the speed of light which Galileo could have measured using this technique?
 - 40 m s^{-1}
 - $4 \times 10^4 \text{ m s}^{-1}$
 - $4 \times 10^6 \text{ m s}^{-1}$
 - $3 \times 10^8 \text{ m s}^{-1}$
- Why was Maxwell so convinced that light was an electromagnetic wave?
 - His equations showed that electromagnetic waves would travel at the same speed as light.
 - His equations showed that electromagnetic waves would affect the path of light.
 - He did many experiments and found that the speed of electromagnetic waves was the same as that of light.
 - He did many experiments and found that light was bent by strong magnetic fields.
- In 1905 Einstein put forward two postulates. Which two of the following best summarise them?
 - All observers will find the speed of light to be the same.
 - In the absence of a force, motion continues with constant velocity.
 - There is no way to detect an absolute zero of velocity.
 - Absolute velocity can only be measured relative to the aether.
- Was Einstein's first postulate inconsistent with the physics of the time?
 - No, it had already been stated by Galileo.
 - No, it was simply an extension of Galileo's principle of relativity to include electromagnetism as well.
 - Yes, it went against Newton's laws.
 - Yes, it was a completely new principle which had not been thought of before Einstein.
- Which one of the following best represents the basis of Einstein's considerations, which eventually led to the theory of special relativity?
 - The results of numerous experiments to determine the speed of light.
 - The work of Isaac Newton and Michael Faraday.
 - His consideration of the consequences of accepting the implications of Maxwell's equations.
 - His own experiments in electromagnetism.
- Which of the following is closest to Einstein's first postulate?
 - Light always travels at $3 \times 10^8 \text{ m s}^{-1}$.
 - There is no way to tell how fast you are going unless you can see what's around you.
 - Velocities can only be measured relative to something else.
 - Absolute velocity is that measured with respect to the Sun.
- Which one or more of the following statements (which are all true in classical physics) was particularly inconsistent with Einstein's second postulate (that the speed of light is the same for all observers)?
 - A net force on an object produces an acceleration.
 - There is no frame of reference in which there is an absolute zero of velocity.
 - The principle of relativity said that the speed of light should depend on the speed of the observer.
 - The principle of relativity said that the speed of light should depend on the speed of the light source.
- 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:
 - could be travelling towards Earth
 - could be travelling towards Alpha Proxima more slowly than you
 - could be stationary between Earth and Alpha Proxima
 - must be heading towards Earth.
- You are in interstellar space and know that your velocity relative to Earth is $4 \times 10^6 \text{ m s}^{-1}$ away from it. You then notice another spacecraft with a velocity, towards you, of $4 \times 10^5 \text{ m s}^{-1}$. Which one or more of the following best describes the velocity of the other craft?
 - Away from Earth at $3.6 \times 10^6 \text{ m s}^{-1}$
 - Towards Earth at $3.6 \times 10^6 \text{ m s}^{-1}$
 - Away from Earth at $4.4 \times 10^6 \text{ m s}^{-1}$
 - Towards Earth at $4.4 \times 10^6 \text{ m s}^{-1}$

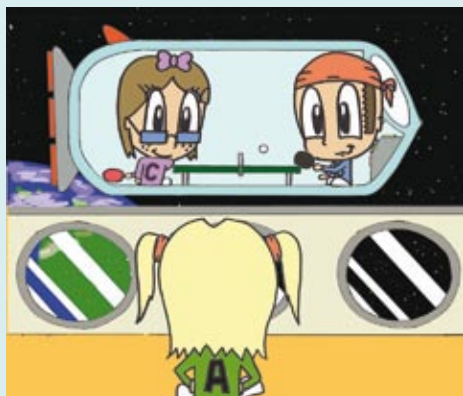
- 11 Spaceships A and B leave the Earth and travel towards Vega, both at a speed of $0.9c$. Observer C back on Earth sees the crews of A and B moving in 'slow motion'. Describe how the crew of A see the crew of B, and how they see C and the Earthlings moving.

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.

- 12 One of the fastest ever objects ever made on Earth was the Galileo Probe which, as a result of Jupiter's huge gravity, entered its atmosphere in 1995 at a speed of nearly $50\,000\text{ m s}^{-1}$. Which of the following is the best estimate of the Lorentz factor for the probe? (You may wish to use the expression $\gamma \approx 1 + v^2/2c^2$.)

A Less than 1
 B 1.000 000 00
 C 1.000 000 01
 D 1.1

The following information applies to questions 13 and 14.



Ben and Chloe are playing table tennis in their space ship as they rush past Anna in her space station at a relative speed of $240\,000\text{ km s}^{-1}$. Ben and Chloe say that they are hitting the ball back and forth with a frequency of 1 hit per second. Their table is 3.0 m long and 1.0 m high.

- 13 Which of the following is the best estimate of the time between hits as measured by Anna?
- A 0.7 s
 B 0.8 s
 C 1.0 s
 D 1.7 s
- 14 How long and high is the table as seen by Anna?
- A 3.0 m long, 1.0 m high
 B 3.0 m long, 0.7 m high
 C 1.8 m long, 1.0 m high
 D 1.8 m long, 0.7 m high
- 15 If a spaceship is travelling at 99% of the speed of light, which of the following best explains why it can't 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 doesn't 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.

- 16 Physicists sometimes say that the mass of an electron is about $8 \times 10^{-14}\text{ J}$. 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 $8 \times 10^{-14}\text{ kg}$.
 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.

Extended-answer questions

In the following questions, assume that phrases such as 'high speed' mean speeds at a significant fraction of the speed of light. In some questions you may need to use the binomial approximation for the Lorentz factor: $\gamma \approx 1 + v^2/2c^2$.

- 17 Make a brief comment on the importance or otherwise, to Einstein's theory, of each of the alternatives A–D in Question 6.
- 18 Make a brief comment on the correctness or otherwise of each of the alternatives A–D in Question 7.
- 19 Aristotle suggested that the 'natural' state of motion for any object is rest. Galileo introduced the principle of inertia which suggested that the natural state of motion is constant velocity (zero velocity being just one example). Explain why Aristotle's view was so hard to shake, and why, if we had spent time as an astronaut in a space station, Galileo's principle would be much easier to accept.
- 20 Maxwell actually felt that there was some sort of error in his equations that predicted the speed of light. Describe the supposed error and the reasons Maxwell was convinced there was a problem with the equations.
- 21 The Michelson–Morley experiment was an attempt to measure the speed of the Earth through the aether.
- a Describe the basic assumption upon which this attempt was based.
- b Why did they use an apparatus which compared the speed of light in perpendicular directions?

- c Briefly describe the results of their experiment.
- 22 An aeroplane can fly at 130 m s^{-1} through the air. The pilot wants to fly to a destination 500 km due north and then fly straight back. However, there is a west wind blowing at 50 m s^{-1} .
- In the absence of any wind, how long would the return trip take?
 - Given the 50 m s^{-1} west wind, how long will a return trip take if the pilot heads the plane so that the actual ground velocity is due north or south?
 - On another occasion, there is a 50 m s^{-1} north wind blowing. Compare the time for a return trip to the same destination on this occasion with that in the part b.
- 23 Compare and contrast the situation in Question 22 and the Michelson–Morley experiment. If the result of the M–M experiment had also applied to the aeroplane in Question 22, what would the pilot have found?
- 24 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?
- 25 Very briefly explain why Einstein said that we must use four-dimensional spacetime to describe events that occur in situations where high speeds and large distances are involved.
- 26 You are in a spaceship travelling at very high speed past a new colony on Mars. Do you notice time going slowly for you; for example, do you find your heart rate is slower than normal? Do the people on Mars appear to be moving normally? Explain your answers.
- 27 Star Xquar is at a distance of 5 light-years from Earth. Space adventurer Raqu heads from Earth towards Xquar at a speed of $0.9c$.
- For those watching from Earth, how long will it take for Raqu to reach Xquar?
 - From Raqu's point of view how long will it take her to reach Xquar?
 - Explain why it is that, although Raqu knew that Xquar was 5 light-years from Earth, and that she was to travel at $0.9c$, it took much less time than might be expected from these figures.
- 28 The space shuttle travels at close to 8000 m s^{-1} . Imagine that as it travels east–west it is to take a photograph of Australia, which is close to 4000 km wide. Because of its speed, the space camera will see everything on Earth slightly contracted.
- About how much less than 4000 km wide will Australia appear to be in this photograph?
 - Will the north–south dimension of Australia be smaller as well?
- 29 As we watch a traveller from Earth to Vega travelling at 99.5% of the speed of light, we will see that their clocks slow down by a factor of about 10 times.
- Explain how this factor of 10 was arrived at.
 - Does this mean that they experience this slowing down of time?
 - Vega is about 25 light-years from Earth, so in our frame of reference it takes light from Vega 25 years to reach us. How long will it take our space traveller to reach Vega?
 - How long will the traveller find that it takes to travel to Vega?
 - Does your answer to part d imply that they were able to get to Vega in less time than light? Explain your answer.
 - If we could travel at the speed of light, how long would it take us to reach Vega?
- 30 The electrons in the Australian synchrotron will be travelling at about 99.9999% of the speed of light. According to our theory of electromagnetism, an electron travelling at this speed in a 1 tesla magnetic field will experience a perpendicular force of $F = qvB = 4.8 \times 10^{-11} \text{ N}$. Now we know that the centripetal force on the electron will be given by $F = mv^2/R$. (The mass of an electron is $9.1 \times 10^{-31} \text{ kg}$.)
- Using this equation, what is the radius of the path of the electron in this magnetic field?
 - Is this answer consistent with what you know about the synchrotron construction?
 - What is the reason for this anomaly?
- 31 In a nuclear power reactor, over the lifetime of the fuel rods about 1 g of every 1 kg of uranium ‘disappears’.
- Has this mass really disappeared?
 - If a reactor was loaded with 1 tonne of uranium fuel, how much energy would be produced from the core of the reactor over the life of that fuel?
 - Melbourne uses electric power at an average rate of about 5 GW ($1 \text{ GW} = 10^9 \text{ W}$). Given that about a third of the energy from the nuclear core will produce electricity, how long would this 1 tonne of uranium fuel power the city?
 - What is the mass of the electrical energy produced by the power station? Why don't we notice this mass?
- 32 The fusion reaction that powers the Sun effectively combines four protons (rest mass $1.673 \times 10^{-27} \text{ kg}$) to form a helium nucleus of two protons and two neutrons (total rest mass $6.645 \times 10^{-27} \text{ kg}$). The total power output of the Sun is a huge $3.9 \times 10^{26} \text{ W}$.
- How much energy is released by each fusion of a helium nucleus?
 - How many helium nuclei are being formed every second in the Sun?
 - How much mass is the Sun losing every day?
 - What happens to this lost mass?

Materials and their use in structures

In 532 AD, 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. It is still a significant structure at 72 m high. In fact, the Hagia Sophia was the tallest building in the world for about 1000 years!

Another remarkable aspect of this building is that it 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. An analysis of the cement that they used has shown that it contained a calcium silicate matrix that is found in many modern cements. 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 have a complete understanding of the properties of the materials with which they work. They must also use their physics knowledge to determine the forces that act within the structures that they have created. This knowledge is the focus of this chapter.

by the end of this chapter

you will have covered material from the study of materials and their use in structures, including:

- the effects of compressive, tensile, shear and bending forces on materials
- the elastic and plastic behaviour of materials
- stress and strain
- strength and toughness
- Young's modulus
- brittle and ductile materials
- strain energy
- torque
- structures in equilibrium.

outcome

On completion of this chapter, you should be able to analyse and explain the properties of construction materials, and evaluate the effects of forces and loads on structures and materials.

7.1

External forces acting on materials

There are a wide variety of materials that exist today and they are used for a wide range of purposes. Some materials are found in nature and others have been manufactured. Over the past 150 years, basalt (or bluestone) has been used extensively in the construction of houses and streets in Melbourne. This material was readily quarried in the area and its properties made it suitable for use in the construction of small houses and laneways. Today, however, bluestone is rarely used for these tasks. It has been replaced by other materials such as bitumen, concrete and steel. Modern materials have properties that enable large structures such as the Eureka Tower and the Bolte Bridge to be constructed.



Figure 7.1 The Bolte Bridge in Melbourne is 490 metres long. It is constructed of prestressed concrete box girders and the longest span in the bridge is 173 m. The steel pile casings that support the bridge are driven 48 m into the ground.

Materials that are used in large structures such as bridges and skyscrapers must be able to withstand enormous forces. The forces acting on a bridge would include the weight of the bridge itself, the weight of the traffic on the bridge, the air pushing as the wind blows and the ground pushing up on the bridge. These forces would cause the parts of the bridge to *stretch*, *compress*, *deform* and *bend* by varying amounts. We will now discuss each of these effects in more detail.



Figure 7.2 (a) These guitar strings are being stretched very strongly and are under tension. (b) This rope is being stretched by its own weight and the weight of the gymnast. The rope is under tension.



Tension

When a stretching force acts on a material, the material is under *tension*. A stretching force is called a *tensile force*. For example, when you stretch a rubber band, you apply a pulling force to each end and place the rubber

band under tension. Most materials do not stretch as noticeably as a rubber band does, but whenever a stretching force is acting, these materials are still under tension. The molecules within the material are pulled further apart.



TENSION is caused by a stretching force and is measured in newtons (N).

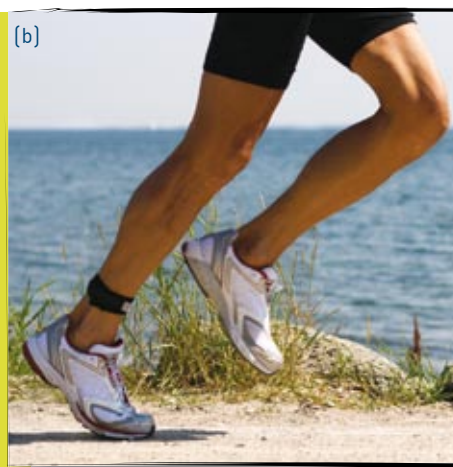
In each guitar string in Figure 7.2a, the tension is equal as you travel along the string. In other words, a string cannot be more taut at one end than at the other. When comparing different strings, however, the tension is most likely to be different from one string to another. The tension in the rope in Figure 7.2b is slightly more complex. The weight of the rope causes the tension to change along the length of the rope: lower tension at the bottom and higher tension at the top. If this rope were to ever snap, it would most likely be at the top of the rope because this part of the rope is holding up all of the rope below it. When the gymnast lets go of the rope, the tensile force would decrease.

Compression

When a material is being squashed or squeezed, it is under *compression*.



COMPRESSION is caused by a squashing force and is measured in newtons (N).



When compressive forces act, the molecules within the material are forced closer together and the material shortens. In the photograph in Figure 7.3a, the pillars supporting the bridge are under compression. They are being squeezed between the ground and the roadway. Similarly, the rubber in the sole of the running shoe in Figure 7.3b is under compression and undergoes a large reduction in thickness.

In the pillars seen in Figure 7.3a, the compressive forces are much larger at the bottom of the pillar than at the top. The material at the bottom is supporting the weight of the pillar above it as well as the roadway. The material at the top is supporting the roadway alone. Pillars such as these are usually made thicker at the bottom so that they can withstand the larger compressive forces acting there.

Physics file

As we stand and walk around each day, our bodies are experiencing forces of compression. Gravity—that is, our weight—is pulling down on us and the force of the ground is pushing upwards on us. At the end of a day, we are slightly shorter than we were at the start because the cartilage in our spine has compressed. The spine then stretches as we sleep and we regain our height. Astronauts who live without the effects of gravity for extended periods grow several centimetres taller during their time in space!

Figure 7.3 (a) The roadway on the Millau Viaduct in France at one point is over 340 m above the ground. The pillars are under compression. They are being squeezed shorter by the downwards force of the roadway and the weight of the pillars themselves and the upwards force from the ground. (b) The materials in the sole of this running shoe are under compression. They are being squeezed between the ground and the athlete's foot each time the shoe lands on the ground.

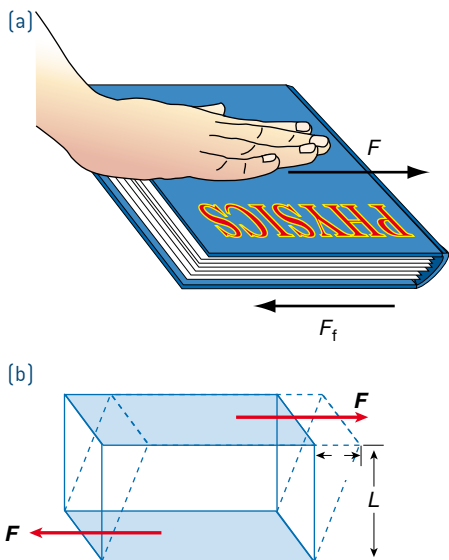


Figure 7.4 [a] As the book is pushed while remaining at rest, an equal and opposite force is supplied by the bench. The resulting shear causes the book to change shape. [b] This diagram shows the effect of the forces on the shape of the book.

Shear

Tension and compression are not the only ways in which forces can act on a material. If a force acts across the top of an object while the bottom remains fixed, a shearing effect or *shear* is said to act.



SHEAR is caused by a pair of sideways forces acting in opposite directions.

To illustrate this, consider a book resting on a table. If a horizontal force is applied across the top of the book, there will be an equal and opposite force acting on the bottom surface in order for the book to remain in equilibrium. The combination of these forces produces a twisting or turning effect within the book, resulting in a shear effect. Typically, an object subjected to shearing forces will not change its linear dimensions significantly, but it will deform or change shape.

Bending

Forces can act on a material in such a way that the material bends. Consider the example of the diver on the end of a diving board. The force that the

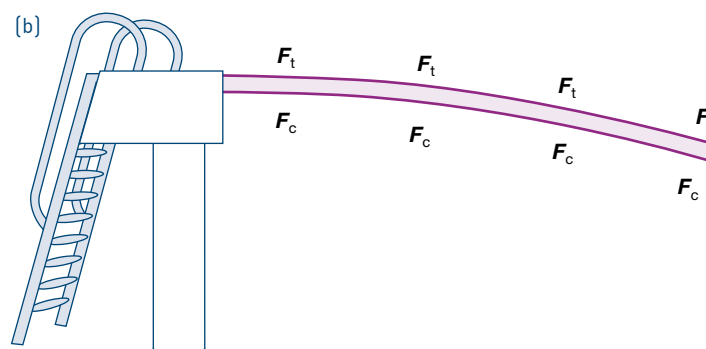


Figure 7.5 [a] The force that the diver exerts on the board makes the board bend downwards. [b] When the diving board bends, tensile and compressive forces act along the top and bottom surfaces, respectively.

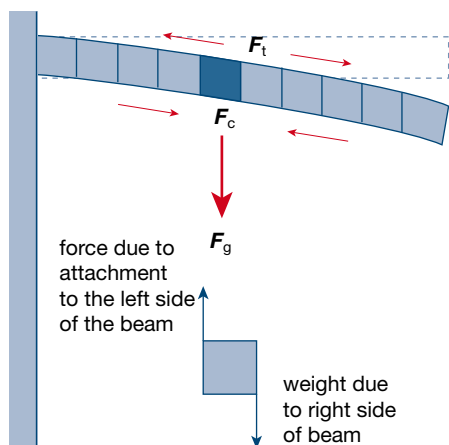


Figure 7.6 The diving board is a beam supported at only one end. This is known as a cantilever. The board will bend under its own weight and bend even more when a diver is on it. Each 'unit cube' of material within the board will experience tensile, compressive and shearing forces.

diver exerts on the board causes the board to bend downwards as shown in Figure 7.5. As the board bends, the material along the top is stretched slightly and so this region is under *tension*. Along the bottom surface, the material is being squeezed slightly shorter and so the board is under *compression* here.

All situations involving bending involve shear. To understand the shear stress in the diving board, think of the board as consisting of many 'unit cubes' (Figure 7.6). One face of each cube will experience a downward force due to the weight of the board while the opposite face experiences an upward force resulting from its attachment to the adjacent cube. This cube is not unlike the book in Figure 7.4, and so experiences a shearing effect.

Forces that affect materials: An overview

Some of the ways in which forces can affect the shape or size of a material are shown in Figure 7.8.

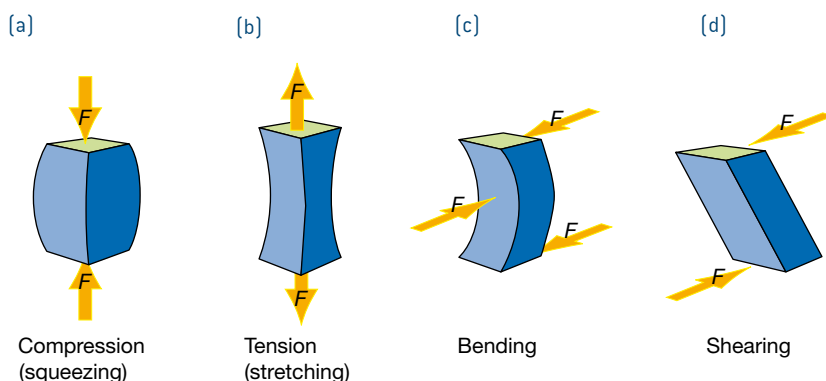


Figure 7.8 The forces that act on or within a material can cause it to (a) be squeezed shorter, (b) be stretched longer, (c) bend and (d) change shape laterally.

Physics file

From time to time tall buildings are subjected to very strong winds. At 200 km h^{-1} (a very strong wind) the force exerted by the moving air is about 1500 N per square metre. On a large building, the force would be enormous. The wind acts to try to push the buildings over and the foundations supply an equal force in the opposite direction, resulting in shear. One feature of tall buildings is their ability to accommodate shearing forces by flexing up to several metres in the wind.



Figure 7.7 The Eureka Tower in Melbourne is 300 m tall. The architects and engineers who designed it would have had to consider the shearing forces that such a tall building would be subjected to. In high winds, the Eureka Tower can flex up to 60 cm . It has two $300\,000$ -litre water tanks on the 90th level to help dampen these oscillations.



7.1 summary

External forces acting on materials

- A material that is being stretched by forces acting on it is under tension. Stretching forces are called tensile forces.
- A material that is being squeezed as a result of forces acting on it is under compression. Squeezing forces are known as compressive forces.
- When parallel, but not colinear, forces act on a material, the material will experience shear. This causes the shape of the material to alter.
- As a material bends, compressive, tensile and shear forces are all acting.

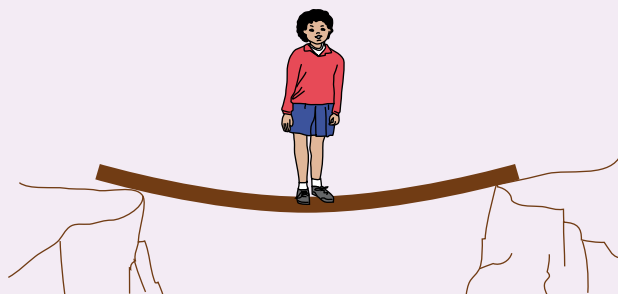


7.1 questions

External forces acting on materials

Use $g = 9.80 \text{ N kg}^{-1}$ when answering these questions.

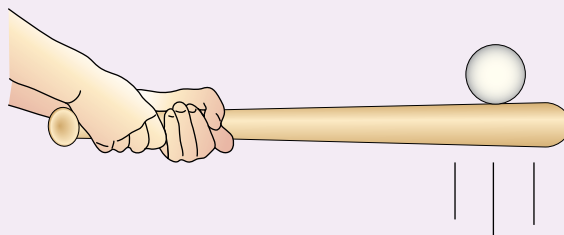
- 1 As a crane is used to lift a concrete slab onto a truck, the cable is experiencing:
 - A compressive forces
 - B tensile forces
 - C shear
 - D tactile forces.
- 2 When a pole-vaulting pole is temporarily bent during a jump, the pole experiences:
 - A compressive forces
 - B tensile forces
 - C shear
 - D all of the above.
- 3 The stumps of a house are usually made of timber or concrete and hold the house up off the ground. What type of force acts within the stumps?
 - A compressive forces
 - B tensile forces
 - C shear
 - D repressive forces
- 4 The wind turbines that have been erected along the coast of Victoria stand about 60 m tall.
 - a Discuss the type of force that acts in the hollow steel towers when the wind is not blowing.
 - b The towers experience enormous shear effects when the wind is blowing. Explain why this is so.
- 5 A girl stands in the middle of a plank of wood that stretches from one side to the other of a 2.0 m wide ditch. Copy the diagram and show clearly where the plank is under compression and where it is under tension.



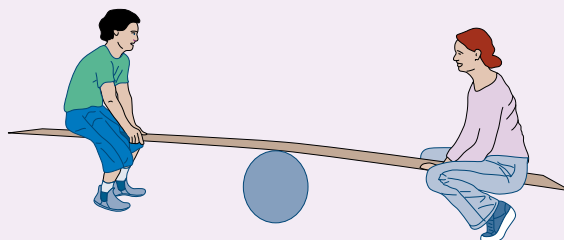
- 6 A boy catches a 3.0 kg fish and holds the fish up while it is still on the hook for his parents to see.
 - a Calculate the tension in the fishing line as the boy stands with his fish.

- b What effect does this force have on the length of the fishing line?

- 7 A tightrope walker is practising her routine above some thick safety mats.
 - a Describe the type of force acting in the tightrope.
 - b She slips off the rope and lands on the mat. What type of force acts in the safety mat as she lands on it?
- 8 During a game of softball, one of the players hits a home run. The diagram shows the bat at the time of impact with the ball.



- a Is the ball under compression or tension at this instant?
 - b Copy the diagram and show the horizontal forces that the player's hands and the ball exert on the bat at this instant.
 - c On the same diagram, show where the bat is under compression and tension as it bends upon striking the ball.
- 9 An 80 kg rock-climber is hanging off the rock face after slipping during the climb. If the climber is at rest and hanging freely on the rope:
 - a calculate the tension in the rope
 - b discuss the effect that the tensile force has on the rope.
 - 10 Shane and his mother are on a seesaw. Copy the diagram and show where the forces of compression and tension are acting in the seesaw as it bends under their weight.



7.2 Stress and strength

As we have seen, when forces act on a material, the material can experience *tension*, *compression*, *shear* or *bending*. The material can change shape as the molecules within it are pulled further apart, squeezed closer together or caused to slide past each other. When these forces act on a material, the material is said to be under *stress*.

If a material is being stretched, it is under *tensile stress*. If a material is being squeezed, it is under *compressive stress*. When shearing forces act, the material will experience *shear stress*.

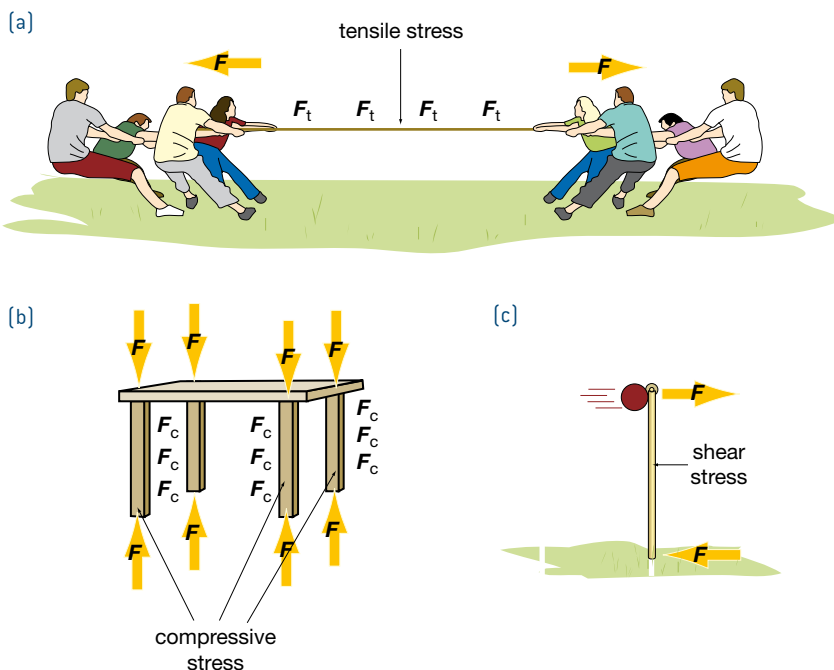


Figure 7.9 (a) This rope is experiencing tensile stress. (b) These table legs are experiencing compressive stress. (c) The cricket stump is under shear stress.

Calculating stress

When a load of force F is applied to a material with cross-sectional area A , in either compression or tension, the material is said to be under *stress*. Stress is defined as the magnitude of the applied force per cross-sectional area of the material. The symbol for stress is σ .

Stress is larger when the applied force is greater. For example, a rubber band is under more stress when you apply greater stretching forces to it. When the stress within a material is greater, the material is more likely to break or *fail*.

Stress also depends on the cross-sectional area of the material. If you apply a 5 N stretching force to a thin rubber band, it will experience greater stress than the same force applied to a thick rubber band. The forces acting in each case are the same, but the thin rubber band is under greater tensile stress because the force is distributed over a smaller cross-sectional area and so the thin band is more likely to break or fail.

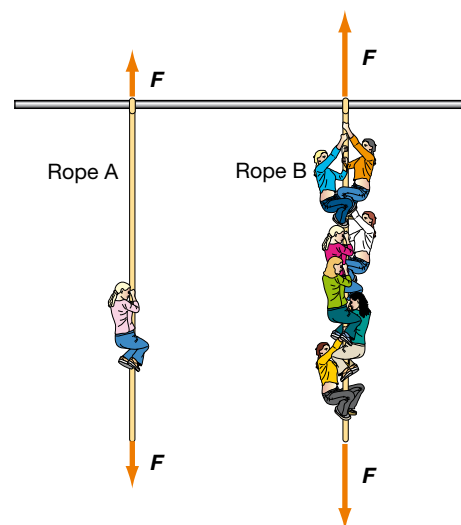


Figure 7.10 These gymnasium ropes are identical, but rope B is under more stress. The forces acting to stretch it are larger, so its tensile stress is greater.

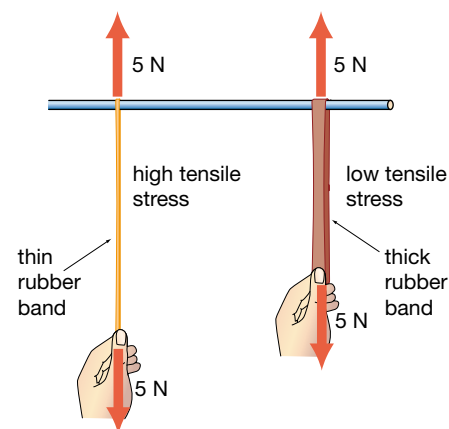
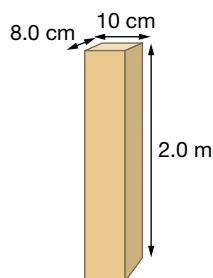


Figure 7.11 The tension in these rubber bands is equal. A 5 N stretching force has been applied to both rubber bands. The stress experienced by each rubber band, however, is not equal. The thin rubber band is under more stress than the thick rubber band because of its smaller cross-sectional area.



Figure 7.12 Nelson's Column in Trafalgar Square, London, provides a useful example of compressive stress. The stress on the material at the bottom of the column is given by $\sigma = F/A$, where A is the area of cross-section for the column and F is the total weight of the column and Nelson's statue (and any pigeons that land on him!). The stress on the material at the top of the column will be less, as the force there is provided only by the statue (and the pigeons).



The **STRESS** (σ) applied to a material is the load force per unit area of cross-section. The load force may place the material under compression or tension:

$$\sigma = \frac{\text{load force}}{\text{cross-sectional area}} = \frac{F}{A}$$

where force is measured in newtons (N)
 area is measured in square metres (m^2)
 stress is measured in newtons per square metre (N m^{-2})

The SI unit for stress is the newton per square metre (N m^{-2}), or pascal (Pa), and typical values can be large. For example, a wire 1 mm in diameter supporting a mass of 10 kg will be under a stress of $1.25 \times 10^6 \text{ N m}^{-2}$. The effect of the stress on the wire is internal. In the unstressed wire, the atoms comprising the wire can be considered to be arranged in planes, each separated by a distance that allows the atoms to be in a position of minimum energy. This means that the attractive and repulsive forces acting on the atoms are in balance. The stress applied as a result of the load places extra tension across the bonds between the atoms, so the bonds are stretched. If the wire were made of steel, the interatomic distance would increase only fractionally. Increases above 1% would be considered large.

Worked example 7.2A

During an experiment, a student hung various weights on the end of some copper wire. In one part of the experiment, copper wire of radius 0.500 mm was used.

- Calculate the cross-sectional area of this wire in square metres.
- Calculate the tensile stress when a load of 3.50 kg was suspended from a length of this wire.
- How great would the stress be if the same load was used on thicker copper wire with a radius of 1.00 mm?

Solution

- The wire has a circular cross-section with radius of $0.500 \times 10^{-3} \text{ m}$ or $5.00 \times 10^{-4} \text{ m}$. Its cross-sectional area is:

$$A = \pi r^2 = \pi \times (5.00 \times 10^{-4})^2 = 7.85 \times 10^{-7} \text{ m}^2$$
- The load force F acting on the wire, assuming g is 9.80 N kg^{-1} , and ignoring the mass of the wire, is:

$$F = mg = 3.5 \times 9.80 = 34.3 \text{ N}$$
 Stress
$$\sigma = \frac{F}{A} = \frac{34.3}{7.85 \times 10^{-7}} = 4.37 \times 10^7 \text{ N m}^{-2}$$
- If the radius of the wire is doubled, its cross-sectional area will quadruple and so the stress will reduce by a factor of four.
 Stress
$$\sigma = (4.37 \times 10^7) \div 4 = 1.09 \times 10^7 \text{ Pa}$$

Worked example 7.2B

A wooden pole has the dimensions shown in the diagram. Four such poles are needed to support an elevated hut to be used by birdwatchers in a rainforest. If the hut has a mass of 250 kg and holds five adults, determine the compressive stress (in MPa) under which the poles are placed. Assume the mass of each adult and their equipment to be 100 kg.

Solution

The force exerted by the hut and birdwatchers is given by the sum of their weights:

$$\begin{aligned}F_g &= mg \\&= [250 + 500] \times 9.80 \\&= 7350 \text{ N down}\end{aligned}$$

The cross-sectional area for each pole is given by $10 \times 8.0 = 80 \text{ cm}^2$, so the total cross-sectional area will be $4 \times 80 = 320 \text{ cm}^2 = 0.032 \text{ m}^2$. The compressive stress will therefore be:

$$\begin{aligned}\sigma &= \frac{F}{A} = \frac{7350}{0.032} \\&= 2.3 \times 10^5 \text{ N m}^{-2} \\&= 0.23 \text{ MPa}\end{aligned}$$

Strength

When engineers assess the properties of a material for a particular task, the two most important considerations are its *stiffness* (How far will it bend?) and its *strength* (How much weight can it support?). Stiffness will be dealt with later in this chapter. Strength can be measured in terms of the *maximum stress* the material will bear before failing. Failure can occur under tension or compression.

The stress at the breaking point under tension is called *tensile strength*. It represents the maximum force that bonds between the layers of atoms can withstand before they are torn apart. For example, if you keep on increasing the stretching force on a rubber band, it will eventually snap or fail. In a similar way, *compressive strength* is the compressive stress at the point of failure under compression. It represents the maximum force that the bonds between the layers of atoms in the material can withstand under compression before the material starts to give way. For example, if you used a metal block to gradually increase the squashing force on an ice cube, eventually the ice cube would shatter under the compressive force. The maximum stress that the ice cube could withstand before it failed is the compressive strength of the ice.



The **STRENGTH** of a material is the maximum stress a material sustains before failing. Strength is measured in pascals (Pa) or newtons per square metre (N m^{-2}).

In theory, both tensile and compressive strength are independent of the shape and cross-sectional area of the material, but in reality different shapes of the same material under compression can support different loads, so engineers have to treat compressive strength values with care.

From Table 7.1, it can be seen which materials are strong under tension and which are strong under compression. Steel is very strong under both tensile and compressive stresses and so is used widely for constructing large buildings and bridges. Timber is not nearly as strong as steel, but timber is still quite robust under both compression and tension. This makes timber suitable for smaller structures such as houses. Granite is much stronger under compression than under tension, which is why it is used in columns but not for cables. Carbon fibre has the opposite properties: it has enormous tensile strength, but no compressive strength. Carbon fibres on their own would not work under compression.



Figure 7.13 Collapsed bridge in Minneapolis, Minnesota, USA, 2007.

Physics file

When designing a structure, architects and engineers must consider the ability of the materials to withstand the forces that will act on the structure. Calculations involving compressive or tensile strengths are invaluable, but they can only be a guide because they are maximum values measured under ideal conditions. The actual material is almost certain to have imperfections in its internal structure, microscopic cracks along its surfaces and the like. A good example involves glass, which is very strong as a thin fibre but in its bulk form is considerably weaker owing to its many microscopic cracks and fissures.

Safety factors, usually from 4 to 10, are a way of accounting for any unknown weaknesses in the materials actually being used. For example, if a steel cable is capable of supporting a load of 500 kg, the manufacturer might consider a safety factor of 5 is appropriate, and so recommend that it be used for masses no greater than 100 kg.

Table 7.1 Typical compressive and tensile strengths for a variety of materials

Material	Compressive strength (MPa)	Tensile strength (MPa)
Cast iron	550	170
Steel	500	820
Brass	250	250
Aluminium	200	200
Timber (pine)		
along the grain	35	40
across the grain	10	
Nylon fibre	0	500
Carbon fibre	0	2000
Marble	80	0
Granite	240	0
Bone	170	130
Concrete	20	2

Worked example 7.2C

A compressive force of $2.50 \times 10^5 \text{ N}$ is applied to a concrete house stump of cross-sectional area 0.012 m^2 . Calculate the compressive stress in the stump and use Table 7.1 to determine whether the stump will undergo structural failure.

Solution

$$\begin{aligned}
 \sigma &= \frac{F}{A} \\
 &= \frac{2.50 \times 10^5}{0.012} \\
 &= 2.1 \times 10^7 \text{ Pa}
 \end{aligned}$$

The compressive strength of concrete given in Table 7.1 is 20 MPa. The pillar has been exposed to greater stress than this value and so will fail.

Composite material: Reinforced concrete

Before the 20th century, most large buildings were made from stone. Stone would be quarried and then carried to the building site, where it was shaped by masons, moved into position, and finally fitted into place. Concrete has replaced stone in almost all construction because it is an inexpensive and convenient building material. It can be mixed at the building site, transported in pipes as a fluid, and then poured into a mould or straight into the building itself. It is durable and very workable.

Concrete, like stone, is very strong under compression, and is ideal for uses such as bridge pylons or the foundations of large buildings. However, Table 7.1 shows that concrete is weak under tension. In fact, its tensile strength is only one-tenth of its compressive strength.

Consider the situation in which concrete is to be used to make the floors in a high-rise building. Under its own weight, a concrete floor will sag at the centre. As it sags, its lower side stretches over a larger area, so the bonds between the atoms will be under tensile stress. At the same time, the upper side will be under compression as the weight of the beam is pulling down

Physics file

Composite materials used in aircraft fuselages weaken over time due to repeated stresses as the aircraft pressurises and depressurises. Scientists are developing 'smart' composite materials that might be able to detect cracks in aircraft bodies before they develop. A composite of epoxy resin and carbon nanotube fibres is sandwiched between a grid of fine wires. By providing a voltage across the wires, the resistance of the composite could be determined. Areas where the fibres had started separating from the resin gave higher resistance readings, indicating early signs of cracking.

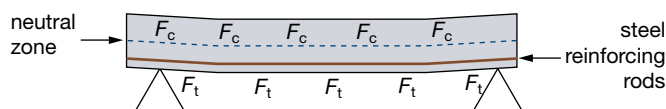


Figure 7.14 A beam of concrete can sag under its own weight. The dotted line represents the neutral zone where the particles in the concrete experience no stress. Above the neutral zone, the concrete is under compression, and below it the concrete is under tension. Where the beam is under tension, small cracks can occur which can significantly weaken the beam. Steel reinforcing is added here.

at the centre. Without support, the tension on the lower side could be great enough to cause small cracks to appear. These cracks may widen without notice, tearing through the whole slab and causing the whole floor to fail. Cracks in a material can travel very quickly—even at speeds of thousands of metres per second!

One way to overcome this problem is to include another material within the concrete that can improve its tensile strength. Steel reinforcing rods with raised ribbing are commonly used. These are laid out and concrete is poured over them, drying around the rods so that the rods are usually near the bottom surface of the beam. The weight of the floor still causes bending and so tension on the bottom surface, but the rods hold the floor together. Reinforced concrete is an example of a *composite material*.

Composite materials consist of two or more materials that retain the properties of their ingredients. A composite material is stronger and/or stiffer than many single materials. Apart from reinforced concrete, there are now a whole host of composite materials available. Fibreglass, developed in the 1960s, is used widely for boat hulls, surfboards, building panels and automotive parts. It consists of thin glass fibres (which are very strong) mixed in a rigid polymer matrix. More recently, even stronger fibres have been used. Boron and carbon fibres are much lighter and stronger than glass and can withstand intense heat. Although these fibres are strong, they are also brittle and susceptible to shear forces. The polymer matrix protects them from damage. Carbon fibres have high tensile strength, but low compressive strength. However, when the fibres are lined up and held in place by the polymer, they exhibit high compressive strength also. These modern composites are widely used in high-impact sports in golf-club shafts and tennis-racquet frames, but are also used in aircraft bodies and engine parts. The fuselages of modern aircraft are often completely made of carbon fibre. Kevlar, a light but very strong polymer matrix, has become a popular material in sailing and is also used in bullet-proof vests.

Composite materials are not necessarily high technology. Mud bricks have been used for centuries as an inexpensive building material. They combine the compressive strength of sun-dried clay with the tensile strength of straw. By itself, the straw is not able to withstand any compressive stress, and dried clay is very brittle, but the mud bricks are very durable.



Physics file

While reinforced concrete can improve the material properties for a concrete slab, the tension that acts along the bottom surface will still cause small cracks to appear. These cracks can let in moisture, which can have devastating—and expensive—results.

An even better solution is to prestress the concrete, causing it to always be under compression and enabling it to support a greater weight. Before the concrete is poured, the ribbed reinforcing steel rods are placed under tension in a frame. After pouring, the concrete sets around the rods and is able to hold them firmly. Once dried, the tension on the rods is released, and as they try to return to their original position, the concrete experiences compressive forces. When a horizontal slab of prestressed concrete is supported at its edges, the lower side remains under compression.

There is no opportunity for tension cracks to appear, and a floor slab made in this way can withstand far greater loads before the lower side experiences any tension at all.

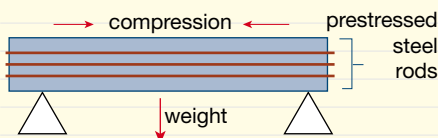


Figure 7.15 A beam of prestressed concrete will always remain under compression, even when lying horizontally.



Figure 7.17 A mud brick is a composite of clay and straw. Mud bricks are used around the world as an inexpensive building material.

Figure 7.16 Steel reinforcing is placed near the lower surface of a concrete slab. The good tensile characteristics of the steel enhance the poor tensile characteristics of the concrete, making the slab more robust. Reinforced concrete is an example of a composite material.



7.2 summary

Stress and strength

- The stress applied to a material is the load force per unit area of cross-section:

$$\sigma = \frac{F}{A}$$

- The load force may place the material under compression (producing a compressive stress) or tension (producing a tensile stress).



7.2 questions

Stress and strength

For the following questions, use $g = 9.80 \text{ N kg}^{-1}$ and make use of the data in Table 7.1.

- Calculate the stress in each of these materials and indicate whether the material is under tensile or compressive stress. Ignore the mass of the material in each case.
 - A climbing rope of radius 0.50 cm supporting a climber of mass 60 kg who is suspended from it
 - A square metal pole of thickness 15 mm that is holding up a loudspeaker of mass 15 kg
- A helicopter used for fighting bushfires is carrying a 900 kg water-filled bucket suspended from a steel cable of diameter 2.4 cm. Ignore the mass of the cable.
 - Calculate the tensile force within the cable when the helicopter is hovering above the fire.
 - Calculate the tensile stress (in MPa) that the cable experiences at this time.

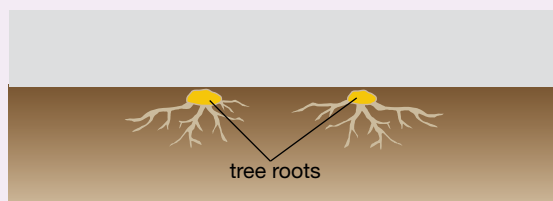
The following information applies to questions 3–5.

A steel rod of length 20.0 cm and radius 5.00 mm is selected at random from a large batch for testing. The rod is placed lengthwise between the jaws of a mechanical vice and a compressive force of 10.0 kN is applied.

- Calculate the compressive stress in the rod.
- What is the maximum compressive force that can be exerted on this rod before it fails?
- The steel rod is to be used as one of the legs of a display stand in a shop. The designer decides that, for safety reasons, the rod should be able to withstand forces five times greater than would be expected under normal working conditions, i.e. a safety factor of 5 is used. What is the maximum compressive force that the designer recommends the rod should be subjected to when used in the display stand?
- Of the 206 bones in the human body, the femur (thigh bone) is the longest. The average human femur has a cross-sectional area of 3.0 cm^2 .

- The strength of a material is the maximum stress that the material can withstand.
- A composite material consists of two or more materials and retains the properties of the materials from which it is made.

- Determine the maximum compressive force that this bone can tolerate before fracturing.
 - Determine the maximum tensile force that this bone can tolerate before fracturing.
- Steel girders are usually produced in the shape of I-beams rather than solid rectangular beams.
 - Explain why not much material is needed along the centre of the beam.
 - Why is the bottom of the beam usually made wider than the top?
 - Marble has a compressive strength of $8.0 \times 10^7 \text{ N m}^{-2}$.
 - Calculate the minimum radius of a cylindrical column of marble that can support a mass of 100 tonnes.
 - What is the minimum radius of a steel column that would support the same load?
 - In a concrete beam laid horizontally across supporting piers, cracks are more likely to appear in the bottom of the beam rather than the top. Explain this fact, using the data in Table 7.1.
 - Explain how the desirable properties of steel and concrete contribute to making reinforced concrete such a useful building material.
 - A concreter is laying a driveway over a section of tree roots. The concreter decides to use steel reinforcing in the concrete to minimise cracking as the tree roots grow. Where in the concrete should the reinforcing be located? Explain your reasoning.



7.3 Strain

When a material is under stress, the bonds holding it together are either tensile or compressive forces. In a structure as large as a bridge or a skyscraper, these forces can be massive. As a result, the materials in these structures will be stretched or squashed to a significant degree. This means that the atom-to-atom distance will be different after a load is applied because the materials are under stress.

For example, if a person were hanging off a length L of bungee rope, it would stretch significantly, as shown in Figure 7.18a. However, if the same person were to hang off a length L of steel cable, the cable would hardly stretch at all. Ignoring the masses of the cable and rope, the tension in both the cable and the rope is the same—given by the weight of the person. The effect of this tensile force on each material is very different. The steel cable is much stiffer and so stretches much less than the bungee rope does. The extent of the distortion of a material is expressed by the *strain*—the fractional change in the length of a material under stress. Strain is represented by the Greek letter epsilon (ϵ) and is always expressed as a positive value, regardless of whether the stress is tensile or compressive. Since strain is the ratio of two lengths, it has no unit. It can be represented as a decimal number (e.g. 0.02) or a percentage (e.g. 2%).



The **STRAIN** (ϵ) is the amount of distortion (i.e. extension or compression) per unit length of the material:

$$\epsilon = \frac{\text{change in length}}{\text{original length}} = \frac{\Delta L}{L}$$

At the atomic level, strain is the factor or percentage by which the bonds between the atoms in the material have lengthened or shortened. Typical values for strain are usually less than 1%, but some plastics can withstand a strain of 5%, and under the right conditions mild steel (steel made using carbon) can be strained by 50%.

To help remember the difference between stress and strain, you may note that **stress** is **pressure** and that **strain** is **length**.

Worked example 7.3A

A rubber sleeping mat is normally 3.5 cm thick. When it is being used, its thickness reduces to 2.0 cm.

- What is the mat's change in thickness?
- Calculate the compressive strain that is acting.
- Express this strain as a percentage.

Solution

- The change in thickness of the mat is 1.5 cm.
- Compressive strain = $\frac{\Delta L}{L}$
 $= \frac{1.5}{3.5}$
 $= 0.43$
- Percentage strain = 0.43×100
 $= 43\%$

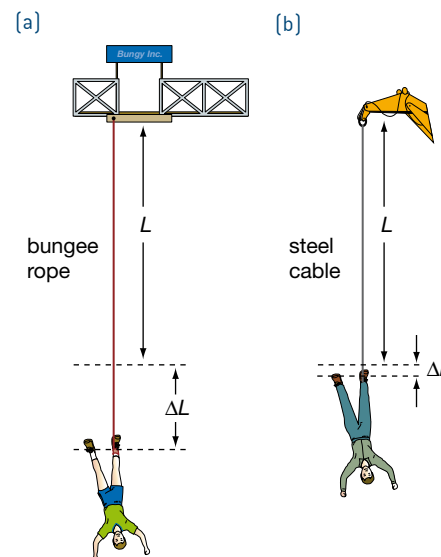


Figure 7.18 (a) The bungee rope changes length by a relatively large amount. The tensile strain is relatively large. (b) The steel cable barely changes length at all. The tensile strain here is very small. The tensile forces in the bungee rope and steel cable are equal.

Physics file

The suspension bridge over the Firth of Forth in Scotland has a 3 km cable that is stretched 3 m by the weight of the road bridge and its traffic. This represents a strain of 0.1%, which is well within safety limits.

Worked example 7.3B

A 1.3 m steel piano wire with a diameter of 1.5 mm is tightened by a force of 900 N, resulting in a strain of 4.3%. Determine the extension of the wire, and calculate the tensile stress on the wire. Comment on whether you think the wire will fail. [Refer back to Table 7.1.]

Solution

$$\text{Since } \epsilon = \frac{\Delta L}{L}$$

$$\begin{aligned}\Delta L &= \epsilon L \\ &= 0.043 \times 1.3 \\ &= 5.59 \times 10^{-2} \text{ m} \\ &= 5.6 \text{ cm}\end{aligned}$$

The tensile stress on the wire will be:

$$\begin{aligned}\sigma &= \frac{F}{A} \\ &= \frac{900}{\pi(7.5 \times 10^{-4})^2} \\ &= 510 \times 10^6 \text{ N m}^{-2} = 510 \text{ MPa}\end{aligned}$$

From Table 7.1, the tensile strength of steel is $500 \times 10^6 \text{ N m}^{-2}$. The wire will probably snap on tightening.



7.3 summary

Strain

- The strain on a material is the amount of distortion (i.e. extension or compression) per unit length of the material:
- Strain is a ratio of two distances and so has no unit. It can be represented as a decimal number or a percentage.

$$\epsilon = \Delta L / L$$



7.3 questions

Strain

Use $g = 9.80 \text{ N kg}^{-1}$ when answering these questions.

- During an experiment, Albert hangs a 1.0 kg mass from a 25 cm long thick rubber band. At the same time, Marie hangs a 1.0 kg mass from a 25 cm long thin rubber band.
 - In which rubber band (thin or thick) is the tension greater?
 - In which rubber band (thin or thick) is the stress greater?
 - In which rubber band (thin or thick) is the strain greater?
- A 20.0 cm steel rod is used as a component in a metal press. During the operation of the press, this rod is subjected to a compressive force of 10.0 kN, under which the rod is compressed by 0.127 mm. Calculate the compressive strain in the rod.
- A piece of timber in a house frame is 4.00 m long and supports a tensile load that produces a strain of 0.128%. What is the elongation of the timber under this load?
- A solid cylindrical steel column of length 7.00 m undergoes a strain of 4.50×10^{-4} while supporting a load. Determine the decrease in length of this column.
- Two marks, 8.000 cm apart, are made on a steel pin that is inserted into a broken bone to assist in the mending process. What will be the distance between these marks when the pin is subjected to a tensile strain of 0.075%?
- A winch uses a steel cable to lift a large piece of machinery from a loading dock. While supporting this load, the length of the steel cable increases to 1.001 times its original value. What is the tensile strain on the cable when supporting this load? Express your answer as a percentage.



- 7 During an interschool hammer throw competition, the wire attached to the hammer is subjected to a tensile strain of 0.002. The unstretched length of this wire is 1.4 m. Calculate the elongation of the wire during the throw.

The following information applies to questions 8–10. During a routine quality control inspection in a precision engineering works, a steel rod of length 20.0 cm and radius 5.00 mm is selected at random from a large batch that has just been completed. This rod is subjected to various tensile forces (F) and the corresponding elongations (ΔL) are recorded in each case. The data obtained are to be displayed in the following table.

F (kN)	ΔL (mm)	σ ($\times 10^7 \text{ N m}^{-2}$)	ϵ (%)
1.0	0.013		
2.0	0.025		
3.0	0.038		
4.0	0.051		
5.0	0.063		

- 8 a Complete this table by determining the values for the stress and strain for the different loads.
b Plot a graph of stress versus strain for this rod.
c Find the gradient for the graph.
d What conclusions can be drawn from this result?
- 9 Use the graph from Question 8b to estimate the strain in a piece of steel that is subjected to a tensile stress of $3.50 \times 10^7 \text{ N m}^{-2}$. Express your answer as a percentage.
- 10 A piece of steel wire with a radius of 1.00 mm is used to support an 80.0 kg chandelier. Use the graph from Question 8b to estimate the tensile strain in the wire while it supports this load.

7.4

Young's modulus

In this section, the concepts of stress and strain are combined to produce a new quantity that provides an exact measure of the *stiffness* and *flexibility* of a material but is independent of its dimensions.

Stress–strain graphs

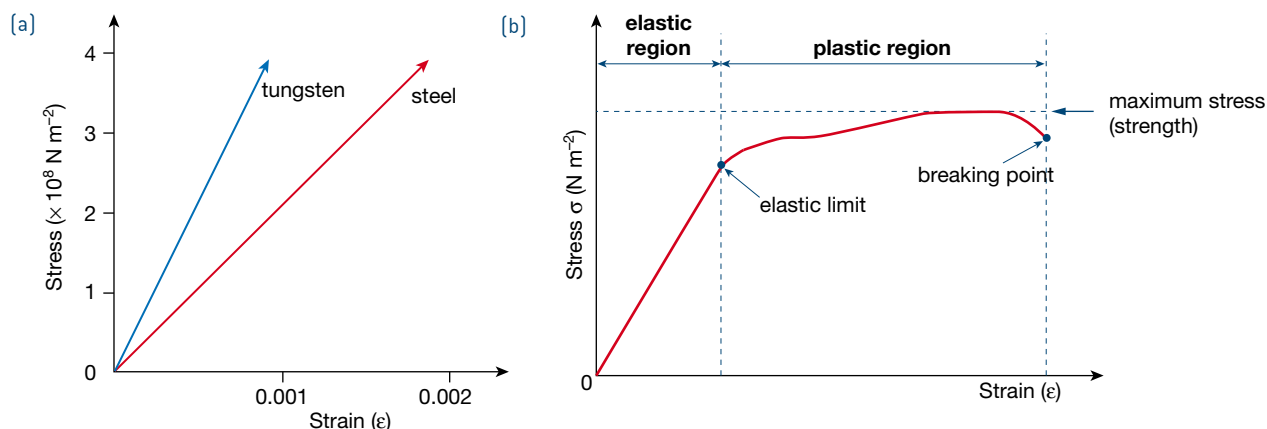


Figure 7.19 (a) Stress and strain plotted for two different materials. Tungsten is stiffer than steel. Steel is more flexible than tungsten. (b) Tensile stress plotted against the corresponding values for the strain for a metal rod that is clamped and subjected to an extending force.



INTERACTIVE TUTORIAL
Young's modulus

Imagine that a metal rod is clamped and its extension is measured as an increasing load is applied. Knowing the cross-sectional area of the rod and its original length, we can calculate both the stress and strain for each load. The stress on the rod is then plotted as a function of the resulting strain, and a graph is produced that is identical in form to the force–extension graph for a spring (as has been studied in Chapter 2). In fact, this should be no surprise. Along the x-axis, the extension of the rod has been replaced by the strain, and since strain equals the extension divided by a constant (the length of the rod) the strain, ϵ , should be proportional to the extension, Δx . Likewise, the stress on the rod equals the applied force divided by another constant (the cross-sectional area of the rod), so the stress, σ , should be proportional to the applied force, F . As these quantities are directly related to each other, the shape of each graph must be the same.

In the linear section of the graph in Figure 7.19b (the *elastic region*), the stress on the rod is directly proportional to the strain, i.e. $\sigma \propto \epsilon$. In this portion of the graph, the material will resume its original dimensions once the stress is removed. This is known as *elastic behaviour*. When a material exhibits elastic behaviour, it returns nearly all of the mechanical energy that went into changing it in the first place. There is very little energy transformed into heat, so the material will *not* heat up significantly.

The *elastic limit* (or yield point) is like the point of no return for the material. Once the stress *exceeds* the elastic limit, the material is *permanently changed* in some way. In this *plastic region* of the graph, the material demonstrates non-linear behaviour. This can be seen in the graph in Figure 7.19b. When the material is in the plastic region, it stretches by a relatively large amount when a relatively small increase in stress is applied. Then when the stress stops acting on the material, the material does *not* return to its original dimensions. It will be longer or shorter depending on whether a tensile or compressive stress was acting. The material has not returned the energy

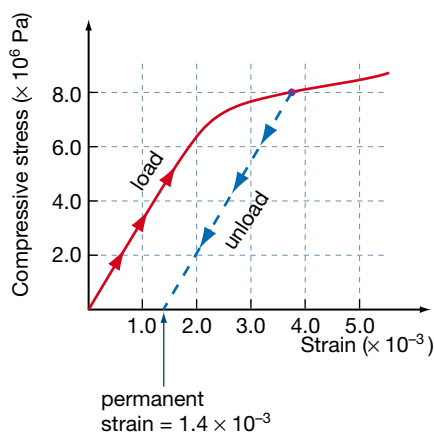


Figure 7.20 If a compressive stress of $8.0 \times 10^6 \text{ Pa}$ is applied to this material, it will undergo plastic behaviour and be permanently strained. This permanent change in length can be determined by first finding the permanent strain as shown on the graph.

that was put into it. Some energy has been used to *change the structure* of the material and some energy is transformed into *heat*. A material that has undergone plastic behaviour will usually heat up significantly.

A stress–strain graph can be used to determine the *permanent strain* produced when a stress in excess of the elastic limit is applied to a material (see Figure 7.20).

Young's modulus

You may recall from studying Hooke's law that the gradient of a force–extension graph for a spring gave a measure of the stiffness of the spring, which was called the *spring constant*. In a similar way, the gradient of a stress–strain graph is a measure of the *stiffness* of a particular material or substance. The gradient is called *Young's modulus* or the elastic modulus of the material. The advantage of using Young's modulus is that it depends only on the material, not on its dimensions. Young's modulus for a length of copper wire is the same as Young's modulus for a copper brick.



YOUNG'S MODULUS (E) for any material is the ratio of stress per unit strain for the material:

$$E = \frac{\text{stress}}{\text{strain}} = \frac{\sigma}{\epsilon}$$

The unit for Young's modulus is N m^{-2} or Pa.

Young's modulus is a measure of the *stiffness* or rigidity of a material. It provides a direct indication of the extent of the distortion that can be expected for a given load. For example, one would think that steel and aluminium were both rigid materials, but Young's modulus shows that steel ($E = 2 \times 10^{11} \text{ N m}^{-2}$) is three times less yielding than aluminium ($E = 7 \times 10^{10} \text{ N m}^{-2}$). A building with an aluminium framework would bend three times farther than an identical building that used steel! Without this measure, comparing the rigidity of any two materials is difficult.

Stiffness should never be confused with strength. Stiffness is measured using Young's modulus, whereas strength comes from the maximum stress that the material can endure before failure. For example, steel is stiff and strong but a wafer biscuit is stiff and weak. In a similar vein, a polymer rope and plasticine are both flexible, but the rope is strong in tension and the plasticine is weak. Substances such as Kevlar, carbon fibre and boron fibre, which are used in modern composite materials, are both stiff and lightweight.

Worked example 7.4A

A nylon string with a diameter of 1.1 mm and an unloaded length of 30 cm is used to replace a broken string in a tennis racquet. When the string is placed under tension, it is stretched by 2.0 cm.

- What is the strain in the string when it is in the racquet?
- Calculate the tensile stress experienced by the string in the racquet.
- What is the force of tension acting on the string in the racquet?
- If the tensile strength for nylon is $5.0 \times 10^8 \text{ N m}^{-2}$, will this nylon string snap?

Table 7.2 Values for Young's modulus for various materials

Material	Young's modulus (N m^{-2})
Carbon fibre	4.1×10^{11}
Boron fibre	4.0×10^{11}
Steel	2.0×10^{11}
Kevlar	1.4×10^{11}
Brass	1.3×10^{11}
Copper	1.2×10^{11}
Glass (crown)	7.1×10^{10}
Aluminium	7.0×10^{10}
Marble	5.0×10^{10}
Granite	4.5×10^{10}
Concrete	2.0×10^{10}
Brick	1.4×10^{10}
Bone (human femur)	1.4×10^{10}
Cast iron	1.0×10^{10}
Timber (pine)	
parallel to the grain	1.0×10^{10}
across the grain	1.0×10^9
Nylon	5.0×10^9
Rubber	4.0×10^6

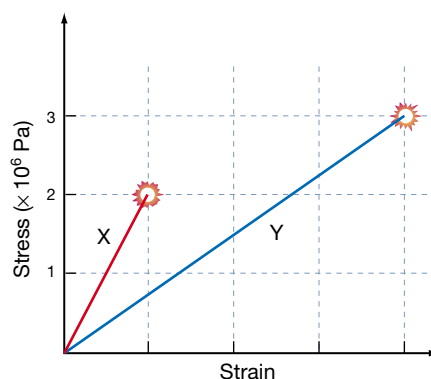


Figure 7.21 Material X is *stiffer* than material Y. X has a greater Young's modulus (gradient) than Y. Material Y is more flexible than X and is also *stronger* than material X. Its maximum stress of $3 \times 10^6 \text{ Pa}$ is greater than the maximum stress that X could withstand ($2 \times 10^6 \text{ Pa}$).

Physics file

Thomas Young was born in Somerset, England, to Quaker parents and was the oldest of 10 children. He became an eminent physicist and physician. In 1801, the young Dr Young performed his double-slit experiment (to be studied in Unit 4) showing the wave properties of light. In his later years, Thomas Young investigated the physiology of the eye and was the first to explain astigmatism, accommodation and colour perception of the retina. He is famously known for his modulus of elasticity relating to materials. Young's modulus depends only on the particular material involved, not its dimensions. This development led to great improvements in engineering strategies.

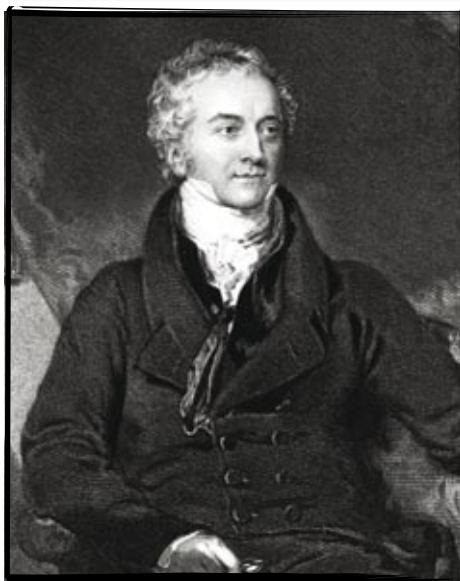


Figure 7.22 Thomas Young (1773–1829).

Physics file

It is important to note that brittle materials are not necessarily weak. For example, some ceramics, which are brittle materials, can have enormous strength. The ceramic tiles used on the space shuttle have a breaking stress much greater than that of most metals.

Figure 7.23 (a) This china plate is brittle. It has not deformed at all. (b) The stress–strain graph for the plate shows that it exhibited no plastic behaviour. Its maximum stress (strength) was also its elastic limit.

Solution

$$\begin{aligned} \text{a } \epsilon &= \frac{\Delta L}{L} \\ &= \frac{2.0}{30} = 0.067 \end{aligned}$$

b From Table 7.2, Young's modulus for nylon is $5.0 \times 10^9 \text{ Pa}$.

$$\begin{aligned} E &= \frac{\text{stress}}{\text{strain}} \\ \Rightarrow \text{stress} &= E \times \text{strain} \\ &= 5.0 \times 10^9 \times 0.067 \\ &= 3.3 \times 10^8 \text{ Pa} \end{aligned}$$

c The cross-sectional area of the string is:

$$\begin{aligned} A &= \pi r^2 \\ &= \pi (0.00055)^2 \\ &= 9.5 \times 10^{-7} \text{ m}^2 \end{aligned}$$

$$\begin{aligned} \text{stress} &= \frac{F}{A} \\ \Rightarrow F &= \text{stress} \times A \\ &= 3.3 \times 10^8 \times 9.5 \times 10^{-7} \\ &= 310 \text{ N} \end{aligned}$$

d In theory, the string should not break as the tensile stress that is acting is less than its tensile strength. However, imperfections in the nylon might cause it to snap.

Brittle and ductile materials

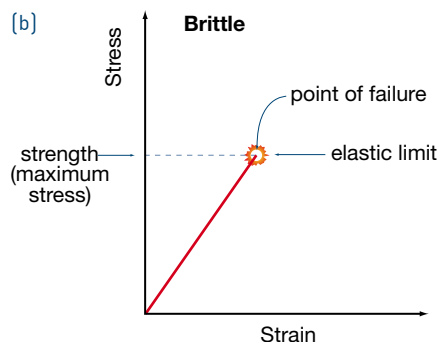
As you will know from experience, objects such as glasses and china plates can smash if you drop them. They can shatter into many pieces. These are known as *brittle* materials. As an increasing stress is applied to these objects, they behave elastically; but once the stress reaches the elastic limit, they fail. The elastic limit of these materials is the same as their maximum stress or strength. They do not demonstrate plastic behaviour and so they do not deform at all. This means that the shattered pieces can often be glued back together after an accident. Even so, you would not want to use brittle materials for making cars. In the event of a car accident, it is preferable that the car remain intact rather than shatter all over the place!

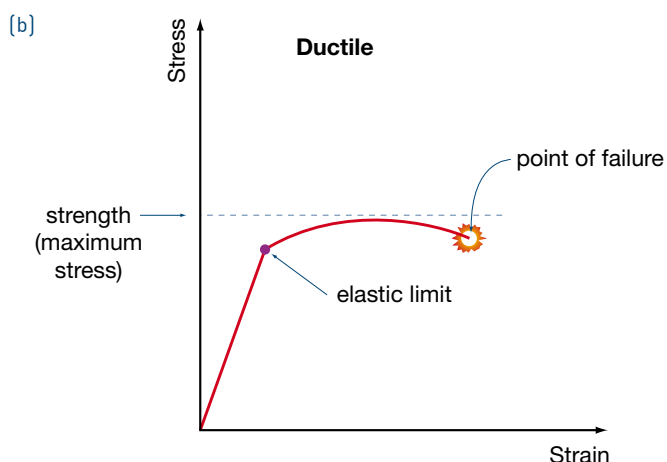
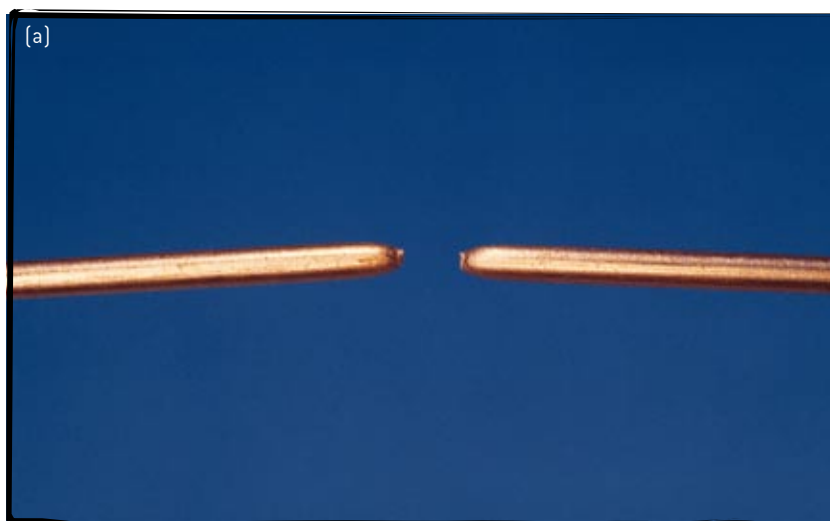
For other materials, in particular metals, the point of failure is usually well beyond the elastic limit. For example, when the stress that is applied to copper wire exceeds the elastic limit, the wire does not fail. It undergoes plastic behaviour and stretches more. The copper is now permanently deformed. Eventually, the copper will fail at a point well past its elastic limit. Materials that behave in this way are called *ductile*. The mechanical energy that has gone into stretching the wire is not conserved. Some is transformed into heat energy and the rest has deformed the material.

(a)



(b)





PRACTICAL ACTIVITY 24

Stress and strain

Physics file

Gold is the most ductile metal. A piece of gold about the size of a small book could be beaten so thin that it could cover a soccer field.

Figure 7.24 Metals such as copper are ductile. This means that they can be drawn into wires. As the copper wire is stretched beyond its elastic limit, one section will narrow and become a weak point. This deformation is called necking. (a) The wire will continue to stretch, then snap or fracture at this weak point. (b) The stress–strain curve for a ductile material shows a large plastic region. The stress at which the copper wire fails is well below the maximum stress that it has withstood.

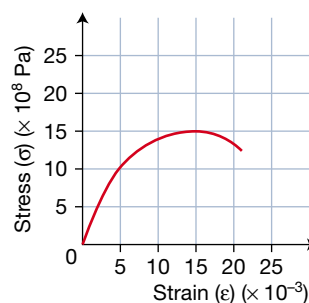
Worked example 7.4B

The stress–strain graph for a steel cable is shown. The force stretching the cable is applied and removed without the cable snapping.

- What is the elastic limit of the material in terms of the applied stress?
- What is the strength of the steel wire?
- After the stretching force stops acting, the cable is permanently stretched. What is its permanent strain (as a percentage)?

Solution

- The cable displays elastic behaviour up to a stress of 10×10^8 Pa, i.e. the elastic limit is 1.0×10^9 Pa.
- The strength is equal to the maximum stress that the material can withstand. From the graph, the strength is 15×10^8 or 1.5×10^9 Pa.
- Draw a line parallel to the elastic region from the end of the stress/strain curve to the strain axis. Permanent strain = $15 \times 10^{-3} = 0.015 = 1.5\%$





7.4 summary

Young's modulus

- Young's modulus, E , for any material is the ratio of stress per unit strain for the material:

$$E = \frac{\text{stress}}{\text{strain}} = \frac{\sigma}{\epsilon}$$

This is given by the gradient of a stress–strain graph.

- Young's modulus provides a measure of stiffness or flexibility for the material that is independent of the dimensions of a particular specimen.
- If a stress lower than the elastic limit is applied, a material will return to its original length and shape when the stress is then removed. When this occurs, the material is said to exhibit elastic behaviour.

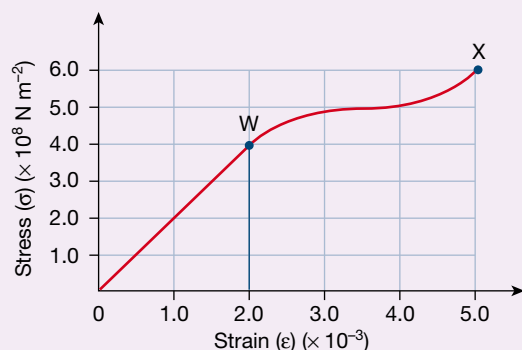
- Stresses greater than the elastic limit may cause a material to undergo plastic deformation. When the stress is removed, the material will not return to its initial dimensions. It remains permanently extended or strained.
- A brittle material can only display elastic behaviour and will fail (break) at its elastic limit.
- A ductile material can withstand stresses greater than its elastic limit and will undergo significant plastic deformation before failing.



7.4 questions

Young's modulus

Use $g = 9.80 \text{ N kg}^{-1}$ when answering these questions. The following information applies to questions 1–4. Steel wire is used to support some extremely heavy pictures during an exhibition at the National Gallery. The following graph shows the stress–strain relationship for the particular steel alloy used. At point W the wire loses its elastic properties, and it snaps at point X.



- a** What is the maximum stress that this wire can tolerate before it undergoes plastic deformation?

b What is the largest strain that this wire will tolerate while still obeying Hooke's law?

c Calculate the gradient of this graph in the interval up to and including the elastic limit.

d What physical constant does the gradient represent?

e What is the tensile strength of this material?
- A 1.0 m piece of this steel wire has a radius of 1.0 mm.

a What is the maximum extension that this wire can tolerate without breaking?

b What is the maximum extension that this wire can tolerate before it undergoes plastic deformation?
- A tensile force of 1.0 kN is applied to the wire described in Question 2. Use the graph above to find the subsequent extension of the wire:

a while the force is still acting

b some time after the force has been removed.
- This wire is now placed under a tensile stress of $5.0 \times 10^8 \text{ Pa}$.

a Choose the correct response. Up to this point, the wire has exhibited brittle/ductile behaviour.

b What is the extension of the wire while it is supporting this load?

c If the crate is removed, will the wire return to its original length? Explain your answer.
- A metal rod 2.00 m long and 1.00 cm^2 in cross-section is subjected to a tensile force of 5.00 kN. As a result its length increases by 0.800 mm. Calculate Young's modulus for the material from which the rod is made.
- a** In general, how does the value of Young's modulus for a particular material relate to its stiffness?

b The value of Young's modulus for three different metals is as follows:

steel $2.0 \times 10^{11} \text{ N m}^{-2}$

aluminium $7.0 \times 10^{10} \text{ N m}^{-2}$

tungsten $3.5 \times 10^{11} \text{ N m}^{-2}$

Rank these metals in order of increasing flexibility.



- 7 The value of Young's modulus for aluminium is $7.0 \times 10^{10} \text{ N m}^{-2}$. An aluminium rod of radius 5.00 mm used in a crankshaft in an engine is subjected to a compressive force during which its length is decreased by 1.0%. Calculate the value of the force acting on the crank.

The following data for human bone applies to questions 8 and 9.

Tensile strength = $1.2 \times 10^8 \text{ N m}^{-2}$

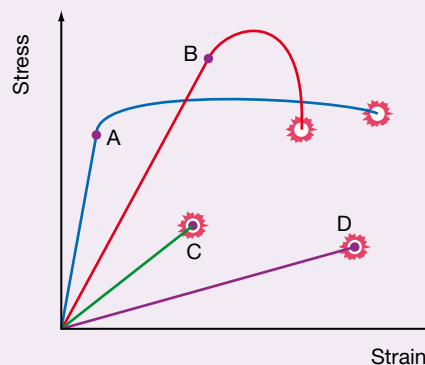
Compressive strength = $1.7 \times 10^8 \text{ N m}^{-2}$

Young's modulus = $1.6 \times 10^{10} \text{ N m}^{-2}$

Elastic limit = $1.0 \times 10^8 \text{ N m}^{-2}$

- 8 The human femur has an average cross-sectional area of 3.0 cm^2 and an unloaded length of 0.40 m.
- What is the maximum compression that this bone can tolerate while behaving elastically?
 - Are bones more likely to break while under tension or compression? Discuss.
 - The bones of elderly people are said to be more brittle than those of younger people. Suggest some reasons why this might be so in the light of your knowledge of materials.
- 9 It is estimated that in sporting activities, the largest compression that a femur of length 40 cm will encounter will be about 0.30 mm. Calculate the compressive stress on the bone for this compression of the femur.

- 10 Use the stress-strain graph to answer the following questions.



- Which one or more of the materials is brittle?
- Which material is the stiffest?
- Which material is the least stiff?
- Which material is the strongest?
- Which material is the most ductile?
- Which line would correspond to a material suitable for building car bodies?

7.5

Strain energy and toughness

In our investigation into the properties of materials so far, much of the discussion has concentrated on the response of materials to a force. In physics, all mechanical situations can also be treated from the point of view of *energy*. When a tall building is subjected to the force of a strong wind, the building will be deflected in response to the force. In the process of deforming the building, the applied force moves the top of the building, so work is done on the building. If the wind suddenly stops (i.e. the load force is removed), some of this work is converted to kinetic energy as the building regains its initial position through its elasticity. But a significant part of the work done will also be transferred to heat energy. Calculating these energies for simple situations will be the focus of this section, but the ideas are most important to those whose job it is to design safe and functional buildings and other structures.

In Area of study 1 'Motion in one and two dimensions', we saw that the area under a force–extension graph gives the *work done* in changing the length of the material, in joules. In a stress–strain graph, the unit for the area under the graph will be different.

Physics file

A simple rubber band does not exhibit simple stress–strain behaviour. Rubber bands do not follow Hooke's law. They behave in a non-linear manner as they are stretched and as they are unloaded before finishing up with close to the original dimensions. The unloading curve is different from the loading curve, indicating that some energy is dissipated during this process. If you stretch and unstretch a rubber band a number of times, you should be able to feel this heat that is lost.

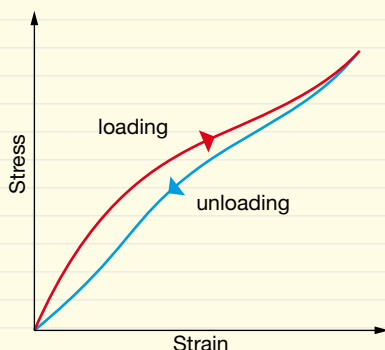


Figure 7.26 The hysteresis curve displayed by a rubber band indicates that it has different loading and unloading behaviour. The amount of energy that is transformed into heat is indicated by the size of the area between the lines.

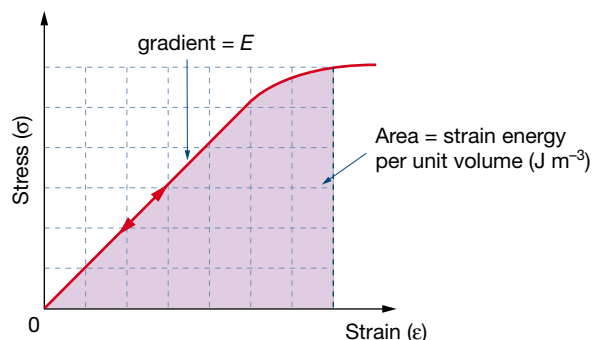


Figure 7.25 The area under the stress–strain graph is the strain energy per unit volume (J m^{-3}) of the material. If the area is an irregular shape, a counting squares technique may be used to estimate its value.

The unit for the area under the graph is newtons per square metre (N m^{-2}), the same as for stress. However, this unit can be transformed into a far more meaningful form. The unit for the area under the graph will equal the unit for the product $\sigma\epsilon$, which is $\frac{F}{A} \times \frac{\Delta L}{L}$ or $\frac{F\Delta L}{AL}$. Since $F\Delta L$ is work done and AL is volume, this is equivalent to work/volume, for which the units are J/m^3 or J m^{-3} .

Thus the *area under a stress–strain graph* gives the *strain energy per cubic metre* or unit volume. The SI unit for this will be joules per cubic metre (J m^{-3}), which can be interpreted as the work done in loading each cubic metre of material. Given this, if the strain energy for a particular sample of material is required, the value of the area under the graph will have to be multiplied by the volume of the material.



STRAIN ENERGY is the work done in changing the length of a material. The strain energy per unit volume of a material can be found from the area under a stress–strain graph.

Strain energy can be determined from a stress–strain graph by multiplying the area under the graph by the volume of the material.

Worked example 7.5A

A glass rod is subjected to a compressive force and undergoes a strain of 0.050%. The stress–strain graph for the glass rod is shown. The rod is 20 cm long and 1.0 cm in diameter.

- Use the graph to determine the strain energy stored in the rod when the strain is 0.05%.
- If the compressive force is removed, will the rod return to its original length?
- Discuss the energy transformations that occur in the glass rod when the force is removed. Assume that it behaves in an ideal manner.

Solution

- A strain of 0.05% represents a strain of 0.0005. From the graph, the corresponding stress is $3.0 \times 10^7 \text{ N m}^{-2}$. The area under the graph is $\frac{1}{2} \text{ base} \times \text{height}$, so:

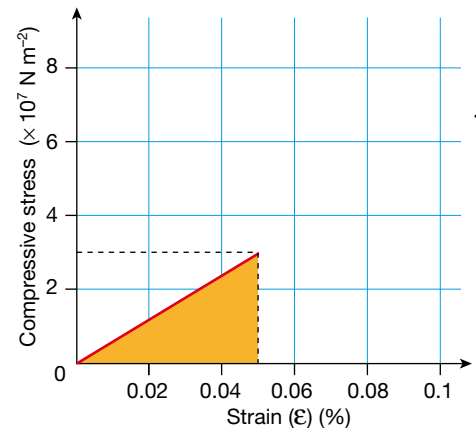
$$\begin{aligned}\text{Area} &= \frac{1}{2} \times 0.0005 \times 3.0 \times 10^7 \\ &= 7.5 \times 10^3 \text{ J m}^{-3}\end{aligned}$$

The volume of the glass will be that of a cylinder ($V = \pi r^2 h$), so:

$$\begin{aligned}V &= \pi \times 0.005^2 \times 0.20 \\ &= 1.57 \times 10^{-5} \text{ m}^3\end{aligned}$$

The energy stored in the glass will be: energy = $7500 \times 1.57 \times 10^{-5} = 0.12 \text{ J}$

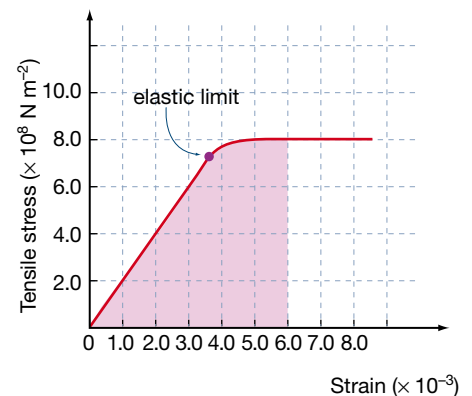
- Yes, the graph indicates that the elastic limit has not been reached so the glass will behave elastically.
- In an ideal elastic material, the strain energy will be completely transformed into kinetic energy with no heating effects in the glass.



Worked example 7.5B

The properties of a 1.5 m length of steel wire of cross-sectional area $8.0 \times 10^{-7} \text{ m}^2$ are shown in the graph on the right. The wire is subjected to a tensile force and experiences a strain of 6.0×10^{-3} .

- By how much has the wire stretched (in mm)?
- Use the stress–strain graph to determine the strain energy stored in the wire when the strain is 6.0×10^{-3} .
- If the tensile force stops acting on the wire, will the wire return to its original length?
- The strain energy that is stored in the wire is not fully transformed into kinetic energy when the tension is removed. How much mechanical energy is lost and where has this energy gone?



Solution

- Strain = $\frac{\Delta L}{L}$

$$\begin{aligned}\Delta L &= 0.0060 \times 1.5 \\ &= 0.0090 \text{ m}\end{aligned}$$

The wire stretches by 9.0 mm

- It is necessary to determine the area under the graph up to a strain of 0.0060. This area is shown on the original graph. Using a counting squares technique gives:

$$\begin{aligned}\text{Strain energy per unit volume} &= \text{area under graph} \\ &= 1.0 \times 10^{-3} \times 2.0 \times 10^8 \times 16 \text{ squares} \\ &= 3.2 \times 10^6 \text{ J m}^{-3}\end{aligned}$$

To find the stored energy, the area under the graph is multiplied by the volume of steel.

$$\begin{aligned}\text{Volume of steel} &= \text{length} \times \text{cross-sectional area} \\ &= 1.5 \times 8.0 \times 10^{-7} = 1.2 \times 10^{-6} \text{ m}^3\end{aligned}$$

$$\begin{aligned}\text{Strain energy} &= \text{graph area} \times \text{volume of steel} \\ &= 3.2 \times 10^6 \times 1.2 \times 10^{-6} = 3.8 \text{ J}\end{aligned}$$

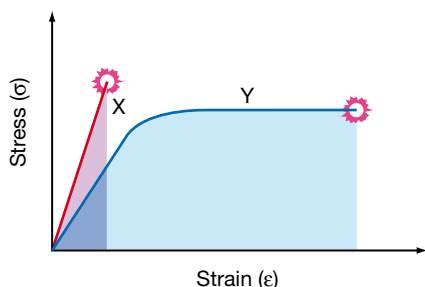
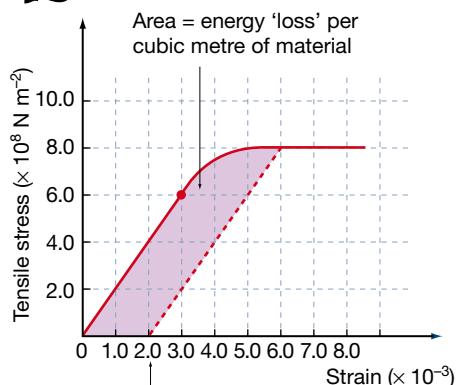


Figure 7.27 Material X is not very tough. While it is very stiff, not much energy is needed to break it, as indicated by the small area under the graph (J m^{-3}). By way of contrast, material Y is very tough. A large amount of strain energy [as indicated by the large graph area] is required before it fails. Y undergoes a lot of plastic deformation before breaking.

Physics file

Most metals are soft and easy to work into different shapes. To make a metal stronger, it is usually combined with other elements to make an *alloy*. Steel is an alloy of iron containing carbon and varying proportions of other metals. Stainless steel contains 12% chromium, which prevents corrosion; nickel increases the strength of steel; and tungsten makes it more heat-resistant. When iron undergoes plastic deformation, the applied forces cause the atoms to slide past each other fairly easily. The additional atoms in steel hold up the movement of the iron atoms, resisting their 'flow'. The result is a stronger material.

- c** No, the elastic limit of the wire has been exceeded.
d When the tension is removed, the wire is permanently strained as shown in the adjacent graph. Some of the energy used to stretch the steel has not been returned. Some of this wasted energy has been transformed into heat energy, causing the wire to heat up and some has been used to permanently change the steel's structure. The area shaded on the graph represents the wasted energy per cubic metre of steel.

$$\text{Shaded area} = 8 \text{ squares} \times 1.0 \times 10^{-3} \times 2.0 \times 10^6 = 1.6 \times 10^6 \text{ J m}^{-3}$$

$$\text{Wasted energy} = \text{shaded area} \times \text{volume of steel} = 1.6 \times 10^6 \times 1.2 \times 10^{-6} = 1.9 \text{ J}$$

Toughness

Most of us have had to carry a polythene bag full of groceries from the supermarket. If the groceries are very heavy, the handles of the bag may begin to stretch. The bag is being stressed beyond its elastic limit, but luckily it can probably undergo significant plastic deformation before failing. This is called *toughness* and it is a very useful property for shopping bags. The deformation is a warning to put the groceries down and rearrange them into other bags before the bag finally gives way. This is preferable to a bag made from paper which, being relatively brittle, will tend to tear without warning. The same is true for larger structures where a tough material may give some warning that it is experiencing too much stress.

Materials that can absorb the greatest strain energy before failure are considered to be the toughest. The toughness of a material is its ability to absorb energy while experiencing plastic deformation. This energy can be estimated by finding the *strain energy* for the material up to the point of failure. Typically, metals are tough, and brittle materials are not. Tough materials also tend to have high Young's modulus values.

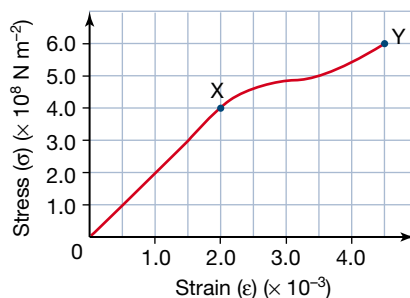


TOUGHNESS is a measure of the amount of energy that a material can absorb before it fails. The toughness of a material is indicated by the area under the stress-strain graph. Tough materials exhibit plastic behaviour.

Worked example 7.5C

The following graph shows the stress-strain relationship for a metallic rod under tensile stress. The elastic limit for this rod occurs at point X, while it fractures at point Y. The rod has a volume of $5.0 \times 10^{-5} \text{ m}^3$.

- a** What is the value of Young's modulus for this rod?
b What is the tensile strength of this rod?
c How much energy per cubic metre of the material can the rod absorb before experiencing failure?
d Calculate the strain energy required to break this rod.



Solution

- a** $E = \text{gradient of elastic region}$

$$= \frac{4 \times 10^8}{2.0 \times 10^{-3}}$$

$$= 2.0 \times 10^{11} \text{ N m}^{-2}$$
- b** From the graph, tensile strength = breaking stress = $6.0 \times 10^8 \text{ N m}^{-2}$
- c** The energy per cubic metre (J m^{-3}) absorbed by the material is given by the area under the graph up to the point of failure. Counting the squares to estimate the area under the graph gives:
 area = $1.0 \times 10^8 \times 0.5 \times 10^{-3} \times 33 \text{ squares}$

$$= 1.7 \times 10^6 \text{ J m}^{-3}$$
- d** The volume of metal is $5.0 \times 10^{-5} \text{ m}^3$
 Strain energy to cause failure = area under graph \times volume of material

$$= 1.7 \times 10^6 \times 5.0 \times 10^{-5}$$

$$= 85 \text{ J}$$

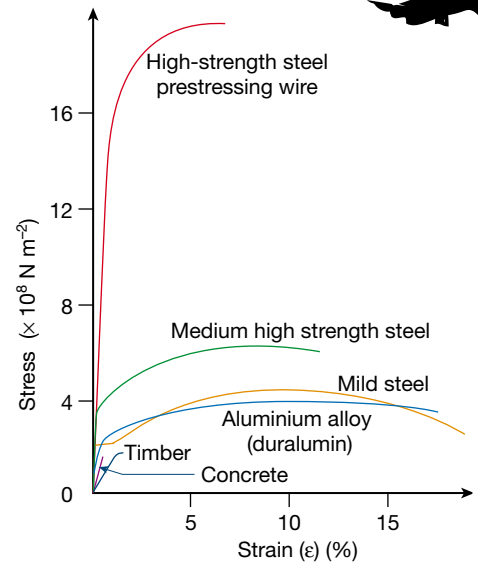


Figure 7.28 The toughness of a material is indicated by the area under its stress–strain graph up to the point of failure. From these graphs you can see that steel is far tougher than aluminium, which is tougher than timber or concrete

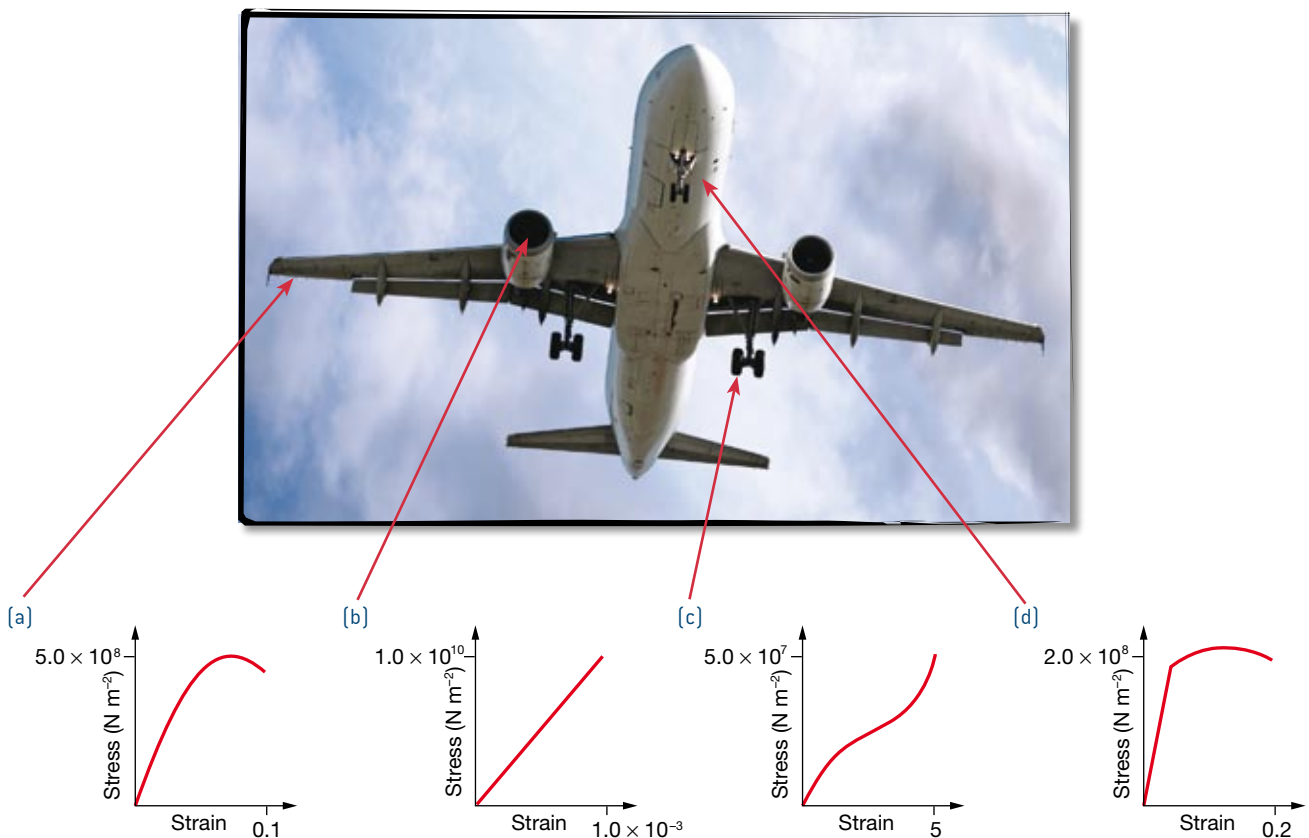


Figure 7.29 Different types of materials that meet different requirements are used in the various parts of an aeroplane. (a) The wings are made of duralumin alloy. This is a stiff material, which is also very tough. (b) The rotor blades are made of a ceramic material, which is enormously strong and which is able to resist high temperatures. (c) The rubber in the tyres is initially stiff, then becomes more flexible before it becomes difficult to stretch again. (d) Copper is used in the electrical circuits. It is ductile, making it easy to draw into wires and it is also an excellent conductor of electricity.

Work hardening and metal fatigue

When a metal is being drawn into a wire, adjacent planes of atoms slide past each other relatively easily. But if the metal has been repeatedly worked or stressed, this will not occur, because many atoms will have been dislodged from their positions of minimum energy. The metal is said to have become *work-hardened*, and is brittle and will snap easily. Try bending a piece of stiff wire back and forth a number of times to demonstrate this phenomenon.

Annealing is the reverse of work-hardening. A material is subjected to a slowly increasing temperature, providing enough thermal energy for the atoms to move back into their positions of minimum energy. The result is a material that can exhibit plastic deformation at very low stress values—which is ideal for changing the shape of the material.

Metals can also suffer from *fatigue*. Continual loading and unloading—even with small stress values—can work-harden small areas of a metal. At these points, the molecular structure is altered and the metal becomes weak and brittle.

The work done by the strain energy will create small cracks. If these cracks enlarge and reach a critical size, they are able to tear through the material at speeds up to many hundreds of metres per second. The material fails—but at stress values far lower than its ultimate strength. It is not surprising that there are stringent safety regulations for structures such as aircraft wings, bridge supports, elevator cables and the like.



Figure 7.30 (a) Before (b) and after! The dramatic collapse of this radio telescope in the United States was caused by metal fatigue. Tiny cracks developed near some bolt holes. The cracks became larger and caused the whole structure to collapse.



7.5 summary

Strain energy and toughness

- Strain energy is the work done in changing the length of a material.
- Toughness is a term that describes the ability of a material to absorb energy before it experiences failure. The greater the strain energy that a material can absorb without failing, the tougher the material. A tough material will undergo significant plastic deformation before failure.
- The strain energy per unit volume (J m^{-3}) of a material is equal to the area beneath the stress-strain graph for the material.
- Strain energy can be found by multiplying the area under a stress-strain graph by the volume of the material.

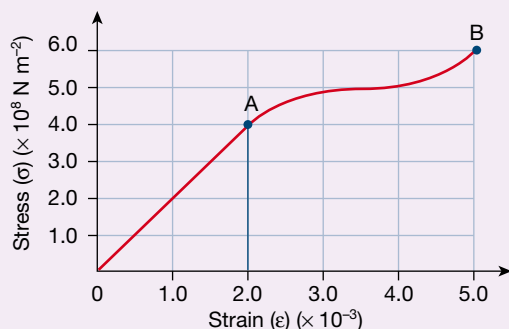


7.5 questions

Strain energy and toughness

The following information applies to questions 1–4.

The stress–strain graph for an alloy being developed by Australian scientists is shown below. The elastic limit is indicated by A, while B indicates the point of fracture. The volume of the alloy is $4.0 \times 10^{-5} \text{ m}^3$.



- The alloy is subjected to a stress and undergoes a strain of 0.0020.
 - Calculate the work done per cubic metre to produce this strain.
 - Calculate the strain energy stored in the alloy at this point.
 - Describe how the alloy will behave when the stress is removed. Will it resume its original length? Will the alloy heat up?
- Now the stress is increased so that the strain experienced by the alloy increases to 0.0040.
 - Calculate the work done per cubic metre to produce this strain.
 - Calculate the strain energy stored in the alloy at this point.
 - Describe how the alloy will behave when the stress is removed. Will it resume its original length? Will the alloy heat up?
- Finally, a stress of $6.0 \times 10^8 \text{ N m}^{-2}$ is applied to the alloy.
 - What happens to the alloy at this point?
 - Calculate the strain energy per cubic metre that was needed to fracture the alloy.
 - Calculate the strain energy needed to cause the alloy to fail.
- Explain the difference between elastic deformation and plastic deformation in terms of energy considerations.

The following data for steel apply to questions 5 and 6.

Young's modulus = $2.00 \times 10^{11} \text{ N m}^{-2}$

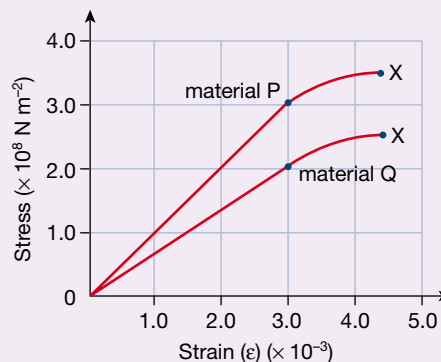
Elastic limit = $4.20 \times 10^8 \text{ N m}^{-2}$

Tensile strength = $8.20 \times 10^8 \text{ N m}^{-2}$

- A cylindrical steel rod of length 50.0 cm and radius 2.00 mm is used in a manufacturing process, where it is subjected to a tensile force of 1.20 kN.
 - Calculate the tensile stress in the rod at this tension.
 - How much strain energy per cubic metre is stored in the rod at this tension?
 - What is the total amount of strain energy stored in the rod under this tension?
- From the data provided, is it possible to obtain an accurate value for the maximum strain energy required to cause this rod to fail? Explain your answer.

The following information applies to questions 7–10.

The stress–strain graphs for two materials P and Q are shown. The points of fracture are indicated by an X.



- When both materials have experienced a strain of 0.30%, calculate the ratio of the strain energy per unit volume of P to the strain energy per unit volume of Q.
- Calculate the minimum energy per unit volume that will cause failure in:
 - material P
 - material Q.
- Which is the tougher material? Justify your answer.
- With reference to the graph, which material is:
 - stiffer?
 - stronger?
 Justify your answers.

7.6

Forces in balance: Translational 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: its weight 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 aircraft's weight downwards, drag backwards (from air friction), 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.

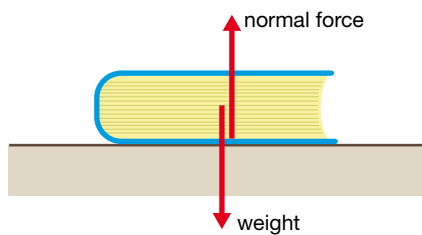


Figure 7.31 A book lying on a table has two forces acting upon it: its weight downwards and a normal force upwards. The net force is zero, so the book remains at rest.

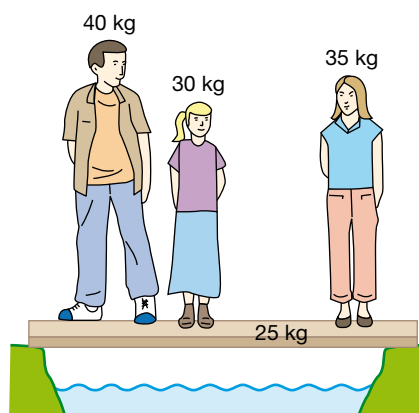


A body is said to be in **TRANSLATIONAL EQUILIBRIUM** when the sum of the forces acting on the body is zero, i.e. $\Sigma \mathbf{F} = 0$.

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 $\Sigma \mathbf{F}_x = 0$ and $\Sigma \mathbf{F}_y = 0$ 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 $\Sigma \mathbf{F}_y = 0$.

Worked example 7.6A

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 kg, 30 kg and 35 kg. Use $g = 9.80 \text{ N kg}^{-1}$ when answering this question.



- Draw a force diagram showing all the forces that are acting on the plank.
- 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 contact with the plank.
- b** The forces acting on the plank are in equilibrium, i.e. $\Sigma F = 0$. This means that the upward forces on the plank are being balanced by the downward forces.

$$F_l + F_r = 392 + 294 + 245 + 343$$

$$700 + F_r = 1274$$

$$F_r = 1274 - 700 = 574$$

$$= 570 \text{ N}$$

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 (weight) 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.

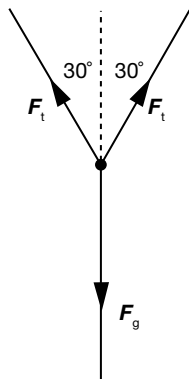
Worked example 7.6B

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.

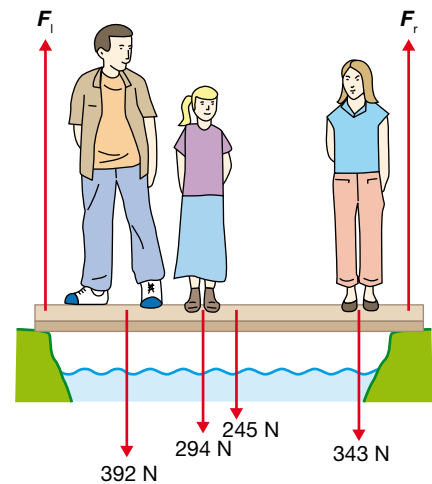
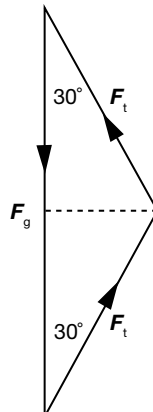
Solution

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 weight, 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).

(a)

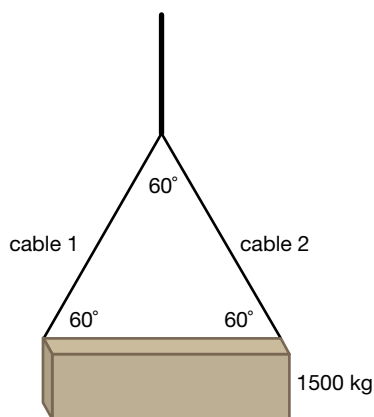


(b)



PRACTICAL ACTIVITY 25

Addition of forces



Using one of the right-angle triangles ($C = A/H$) in (b) gives:

$$H = \frac{A}{C}$$

$$\text{so } F_t = \frac{0.5F_g}{\cos 30^\circ} = \frac{0.5 \times 45 \times 9.80}{\cos 30^\circ}$$

$$= 254.6 \text{ N} = 250 \text{ N}$$

The tensile force is 250 N.

Worked example 7.6C

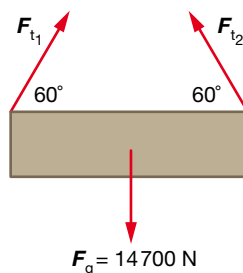
A concrete beam of mass 1500 kg is being lifted by steel cables, as shown in the diagram. The beam is moving upwards with a constant velocity of 2.0 m s^{-1} . Ignore the mass of the cable and use $g = 9.80 \text{ N kg}^{-1}$ when answering the following questions.

- Draw a force diagram showing all the forces acting on the beam.
- Calculate the tension in cable 1 and cable 2 [in kN].

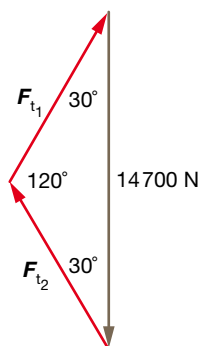
Solution

- There are three forces acting on the beam: the tensile forces acting in each cable F_{t_1} and F_{t_2} as well as the weight F_g of the beam. The tensile forces act at each end of the beam and the weight force acts at the centre of mass.

(a)



(b)



- The beam is moving with a constant velocity, so the forces acting on it are in equilibrium, i.e. $\Sigma F = 0$. Adding the three forces together as vectors (shown above) gives an isosceles triangle. The sine rule can be used to solve this (or the triangle can be broken up into right-angled triangles):

$$\frac{F_{t_1}}{\sin 30^\circ} = \frac{14\,700}{\sin 120^\circ}$$

$$F_{t_1} = F_{t_2} = 8490 \text{ N}$$

The tensile force in each cable is 8.5 kN.



7.6 summary

Forces in balance: Translational equilibrium

- A body is in translational equilibrium when the sum of the forces acting on it is zero, i.e. $\Sigma \mathbf{F} = 0$.
- 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. $\Sigma F_x = 0$, $\Sigma F_y = 0$ and $\Sigma F_z = 0$.



7.6 questions

Forces in balance: Translational equilibrium

Use $g = 9.80 \text{ N kg}^{-1}$ when answering these questions.

- 1 Which of the following are in translational equilibrium?

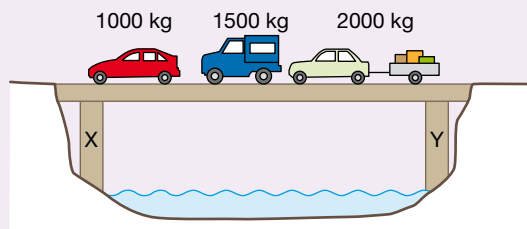
- A A stationary elevator
- B An elevator going up with constant velocity
- C An aeroplane during take-off
- D A container ship sailing with constant velocity
- E A car plummeting off a cliff

- 2 Estimate the tension in a single vertical cable which supports:

- a yourself
- b a small car
- c a full 10-litre mop bucket.

- 3 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.

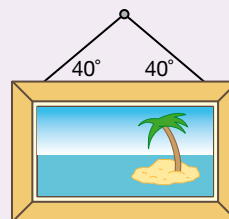
- 4 Three cars are crossing a beam bridge of mass 500 kg in a single line. At one instant, the left pillar (X) is providing a force of $2.0 \times 10^4 \text{ N}$ upwards. What is the size and direction of the force exerted by pillar Y at this instant?



- 5 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?

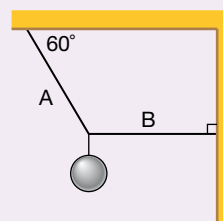
- 6 A rectangular advertising sign is supported from its upper corners by two cables, each making an angle of 40° to the vertical. The sign has a mass of 5.0 kg. Calculate the tension in each cable.

- 7 A picture is hung as shown in the following diagram. If the hanging wire has a breaking strength of 40 N, what is the maximum mass of the picture that can be supported before the wire snaps?



- 8 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.

- 9 A 100 kg plaster ornament is supported by two steel cables of cross-sectional area 1.5 mm^2 . Cable A is at an angle of 60° with the ceiling, and cable B is perpendicular to the wall. The tensile strength of the steel used in the cables is 820 MPa.



- a Assuming the mass of each cable is negligible, calculate the tension in each of them.
- b Which of the two cables is more likely to break?
- c Calculate the stress (in MPa) in each cable.
- d Will either of the cables fail under this load? Explain.

- 10 The arrangement of the ornament in Question 9 is altered so that the angle between cable A and the ceiling is now 30° . Cable B is still perpendicular to the wall. Will either cable fail in this new arrangement? Explain.

7.7

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 7.32 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 point at which the force acts. The amount of torque created is directly proportional to the distance between the axis of rotation and the point at which the force is applied. This distance is called the lever arm and is given the symbol r
 - the angle, θ , between the force and 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: the maximum effect will be achieved by applying a force to the door as far from the hinge as possible. 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. If this angle is reduced, a

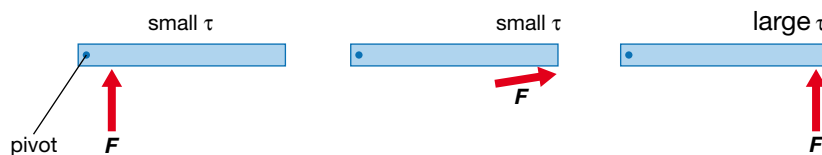


Figure 7.33 The size of a torque depends on the distance of the force from the axis of rotation and the angle that the force makes with the lever arm.

smaller component of the force is perpendicular to the door 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), the door will not rotate, no matter how great a force you apply.

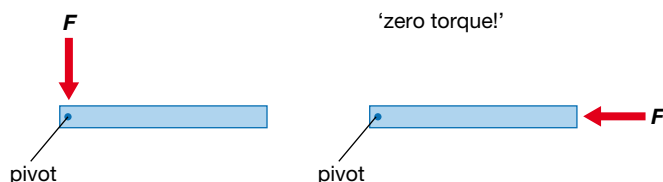


Figure 7.34 No torque is created in these situations. The forces do not generate any rotational or turning effect.



The **TORQUE** (τ) acting on a body is given by the product of the component of the applied force acting perpendicularly to the lever arm, F_\perp , and the distance from the axis of rotation to the applied force, r :

$$\tau = rF_\perp = rF\sin\theta$$

The unit for torque is the newton metre (N m).

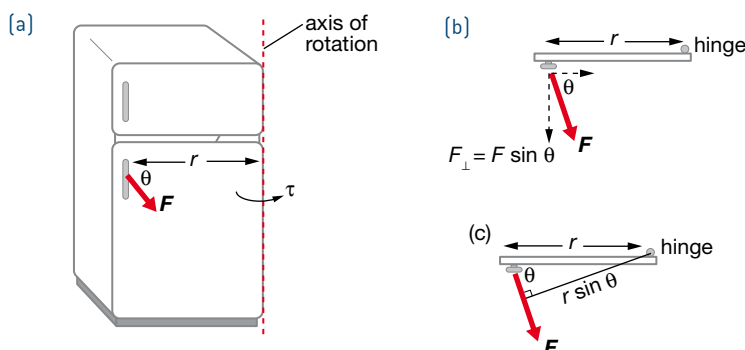


Figure 7.35 (a) A torque applied to a door acts around the axis of rotation (the hinges). (b) This torque is supplied by a force with a component $F_\perp = F\sin\theta$ acting perpendicular to the door at a distance r from the axis of rotation, i.e. $\tau = rF_\perp$. (c) The torque is also the product of the applied force F and perpendicular distance from the force to the hinges, i.e. $\tau = r_\perp F$. Either way, $\tau = rF\sin\theta$.

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.

Worked example 7.7A

A woman whose car has a flat tyre has two wheel-nut spanners in the boot of her car. One wheel spanner is 15 cm long and the other is 75 cm long.

- In order to undo the wheel nuts with a minimum amount of effort, which wheel spanner should the woman select?
- If the maximum force that the woman can apply is 45 N, determine the maximum torque that can be delivered to a wheel nut.

Solution

- 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 $\tau = rF_\perp$, 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!
- Maximum torque will be obtained by applying the 45 N force perpendicular to the lever arm at the maximum distance of 75 cm from the wheelnut. Thus:

$$\begin{aligned}\tau &= rF_\perp \\ &= 0.75 \times 45 \\ &= 34 \text{ N m}\end{aligned}$$

Physics file

Although the unit for torque (N m) appears to be the same as that for work ($1 \text{ J} = 1 \text{ N m}$), 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. Here the force has a component in the direction of the motion. Torque ($\tau = rF_\perp$) is the product of a force and the distance at which that force is acting from a given point. Here, the force acts in the direction of the rotation.

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 the displacement and velocity at any time can be calculated using the principles discussed in Area of study 1 'Motion in one and two dimensions'. 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.

In designing structures, engineers and architects want to ensure that balance and stability are maintained. Whether this occurs depends 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 7.37a 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.

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

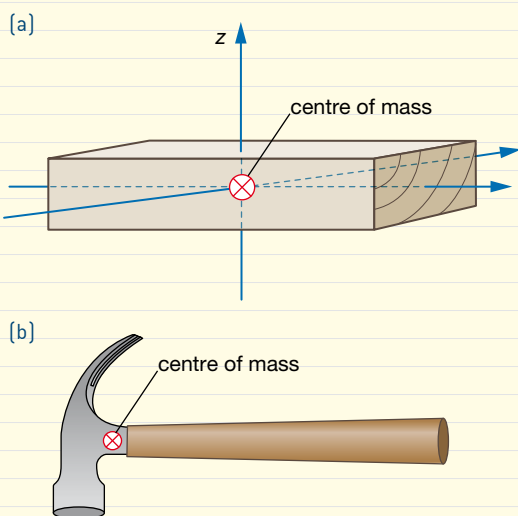
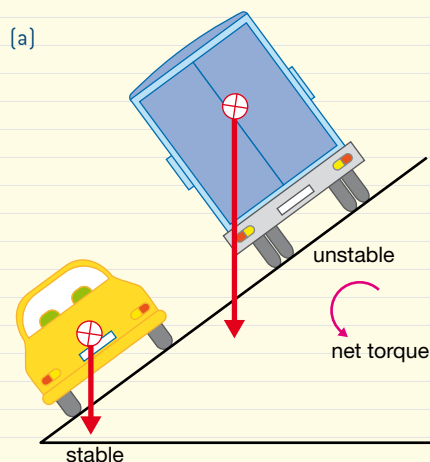


Figure 7.36 The centre of mass for: (a) a uniform body in three dimensions and (b) a hammer.

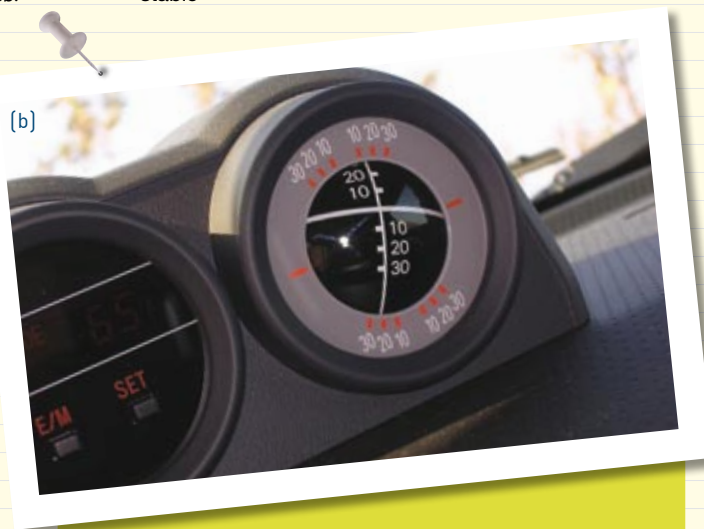


Figure 7.37 (a) The car on the incline is in stable equilibrium, while the heavily laden truck on the same incline could topple. The weight vector is outside the lower point of support for the truck, so there is no reaction force from the road to the higher wheel. (b) Modern four-wheel drives and tractors have inclinometers to warn the driver if the vehicle is in danger of tipping.

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 7.38 The owner of this cart did not have a good understanding of stability and balanced torques!



7.7 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, F , perpendicular to the lever arm and the distance, r , to the axis of rotation:

$$\tau = rF_{\perp} = rF \sin \theta$$
- A force acting directly towards or away from the pivot point produces no torque.



7.7 questions

Torque

Use $g = 9.80 \text{ N kg}^{-1}$ when answering these questions.

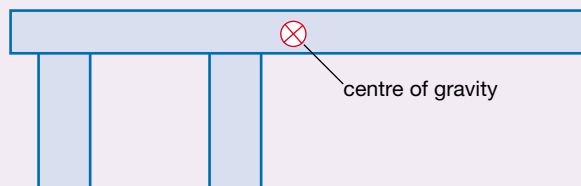
- In the following situations, a torque is acting. In each case, identify the axis of rotation or pivot point about which the torque acts and also estimate the length of the lever arm.
 - A garden tap is turned on.
 - A wheelbarrow is lifted by the handles.
 - An object is picked up with a pair of tweezers.
 - A screwdriver is used to lever open a tin of paint.
- Use the concept of torque to explain the following.
 - It is easier to open a heavy door by pushing it at the handle rather than in the middle of the door.
 - It is possible to move very heavy rocks in the garden by using a long crowbar.
- Renae and Stefan, each of mass 40 kg, are sitting at opposite ends of a playground see-saw. The see-saw is stationary in the horizontal position and is 4.5 m long. Stefan decides to jump off. Calculate the size of the unbalanced torque that now acts on the see-saw.
- A wheelbarrow measures 1.6 m from the tips of the handles to the wheel axle. The contents of the wheelbarrow produce a clockwise torque of 400 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° to the lever arm.
- A crane with a horizontal lever arm is lifting a concrete wall of mass 2.5 tonnes. The load is 20 m from the axis of rotation.
 - Calculate the torque created by this load.
 - What stops the crane from toppling over as a result of this torque?

6 Nikki is investigating torque using a metre rule and a 1.0 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 cm mark.

- What is the size of the torque that is acting?
- She now moves the mass so that it is right at the far end of the ruler. How much torque is acting now?
- Finally, she lifts the ruler so that it makes an angle of 60° 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.

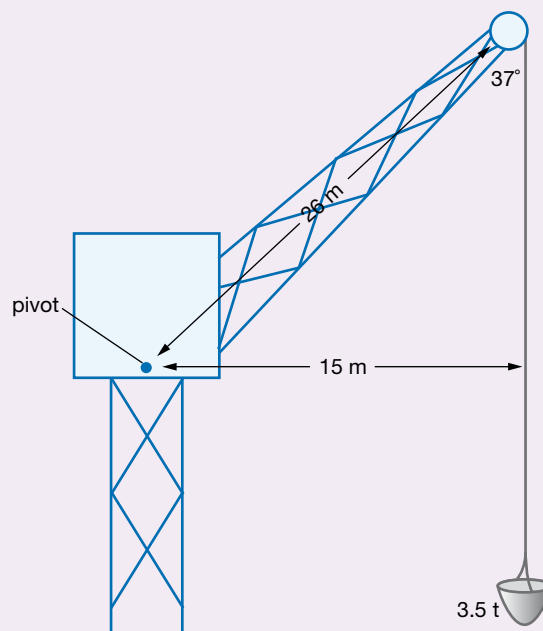
8 Christopher likes constructing things, but he does not have a good understanding of stability. He draws a design for a garden bench that consists of a long beam resting on top of two supports, as shown in the diagram. The centre of gravity of the beam is indicated. Christopher does not intend to use 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 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 3.5 tonnes. The lever arm of the crane is 25 m long and makes an angle of 37° with the vertical as shown in the diagram. Ignore the mass of the cable when answering these questions.



- What is the total weight of the skip?
- 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?
- Calculate the magnitude and direction of the torque about the pivot that the skip exerts on the crane when the skip is at the highest point.

7.8 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. $\Sigma \mathbf{F} = 0$) 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.

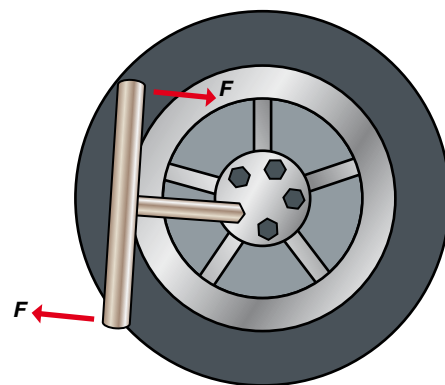


Figure 7.39 Equal and opposite forces are acting on the wheel spanner, but it is not in equilibrium.

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:

$$\Sigma \tau = 0$$



Figure 7.40 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.

Static equilibrium

When a body or system is not accelerating or rotating, it is in *both* translational and rotational equilibrium (i.e. $\Sigma \mathbf{F} = 0$ 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 7.8A

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.

- What is the force supplied to the plank by the pivot when both children are sitting on the plank?
- Where must the girl sit in order to balance the boy?

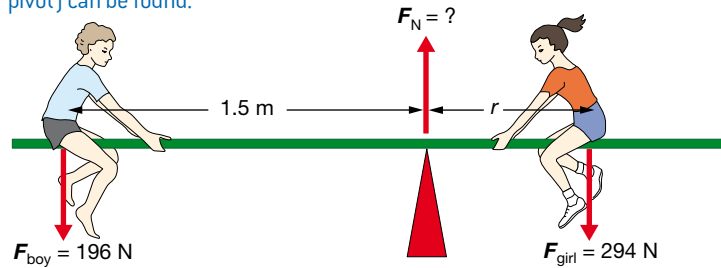


PRACTICAL ACTIVITY 26

Forces in equilibrium

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 normal force supplied by the pivot) can be found.



$$\Sigma F = F_g(\text{boy}) + F_g(\text{girl}) + F_N = 0$$

We let the downward direction be negative, so

$$-[20 \times 9.8] - [30 \times 9.8] + F_N = 0$$

$$-196 - 294 + F_N = 0$$

$$F_N = +490 \text{ N} = 490 \text{ N upwards}$$

- 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.

$$\text{So, } \tau_{\text{boy}} = \tau_{\text{girl}}$$

$$\text{Hence, } 196 \times 1.5 = 294 \times r$$

$$\text{So that } r = \frac{1.5 \times 196}{294} = 1.0 \text{ m}$$

This could also be solved by assigning a negative sign to the clockwise torques and a positive sign to the anticlockwise torques, and having all the torques added to give zero.

In Worked example 7.8A, 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 7.41). This will mean that $\tau_{\text{girl}} = 0$, since the lever arm here will be zero.

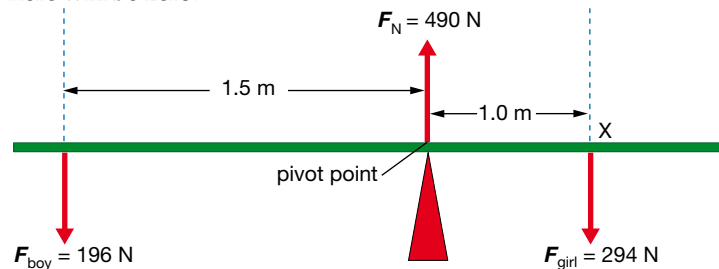


Figure 7.41 A force diagram for the see-saw problem, where the point at which the girl sits (labelled X) has been chosen as the reference point.

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 $\tau_{\text{boy}} = rF = 2.5 \times 196 = 490 \text{ N m}$ anticlockwise, while the torque caused by the normal force will be $\tau_N = rF = 1.0 \times 490 = 490 \text{ N m}$ 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 $r = 0$) and the relationship resulting from the condition $\Sigma\tau = 0$ will solve the other unknown force. Worked example 7.8B employs this strategy.

Worked example 7.8B

While painting a tall building, a 70.0 kg painter stands 4.00 m from the 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 force diagram for the plank, which is in static equilibrium. Here, $\Sigma F = 0$ and $\Sigma\tau = 0$.

Four forces are acting on the plank: the weight of the plank, the weight of the painter pushing down on the plank, and the tension in each of the two ropes. The weight of 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 (F_{t_1} and 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.

$$\tau_{\text{clockwise}} = \tau_{\text{anticlockwise}}$$

$$(3.00 \times 196) + (4.00 \times 686) = 6.00 \times F_{t_2}$$

Rearranging:

$$F_{t_2} = \frac{588 + 2744}{6.00} = 555 \text{ N upwards}$$

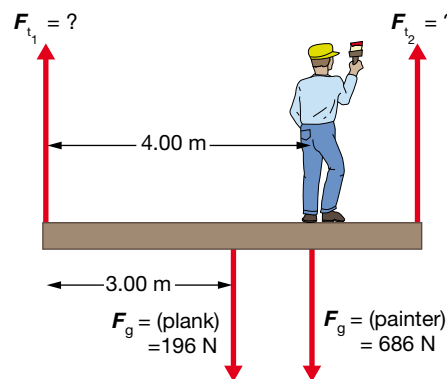
Similarly, the tension in the left-hand rope can be obtained by shifting the reference point to any other position. (We will use the right end of the beam here.) This time, the plank and the painter provide anticlockwise torques and the rope supplies a clockwise torque.

$$6.00 \times F_{t_1} = (3.00 \times 196) + (2.00 \times 686)$$

Rearranging:

$$F_{t_1} = \frac{588 + 1372}{6.00} = 327 \text{ N upwards}$$

To check these values: if $\Sigma F = 0$ then the sum of the two upward forces (tensions), $555 \text{ N} + 327 \text{ N} = 882 \text{ N}$, will balance the sum of the two downward forces, $196 + 686 = 882 \text{ N}$, 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

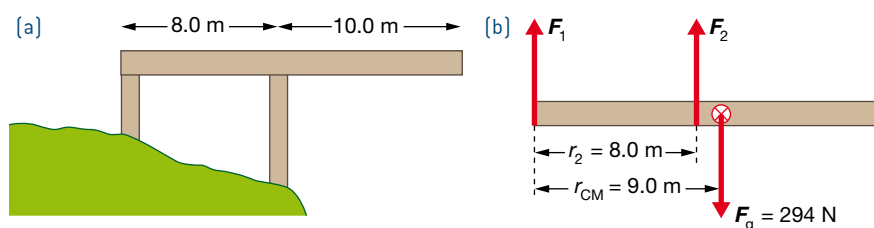


Figure 7.42 (a) The bowsprit projecting from the prow of a sailing ship is a cantilever, and the masts have many vertical cantilevers. (b) The Westgate Bridge is a cable-stayed girder bridge spanning the lower Yarra River in Melbourne. It collapsed during construction in 1970 when a 112 m long cantilevered beam was being positioned. Tragically, 35 workers lost their lives. The bridge eventually opened in 1978.

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 Melbourne, the Museum entrance, Exhibition Centre and Citylink gateway all feature cantilevered designs.

Worked example 7.8C

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, and so its centre of gravity is 9.0 m from each end, to the right of the central pillar. Not knowing in which direction the forces act, we will assume that both support forces, F_1 and F_2 , act upwards. Because of this, we will need to be very careful when using signs for both torques and forces. We will let upward forces be positive and anticlockwise torques be positive. A force diagram shows the location and direction of all the forces. 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 F_1 is acting, then $\Sigma\tau = 0$.

$$(r_2 \times F_2) - (r_{CM} \times F_g) = 0$$

$$(8.0 \times F_2) - (9.0 \times 294) = 0$$

$$\Rightarrow F_2 = 330.75 \text{ N, or } 330 \text{ N upwards}$$

Rather than using the equilibrium condition for torque again, we can find F_1 using the condition for translational equilibrium. If $\Sigma F = 0$, then:

$$F_1 + F_2 + F_g = 0, \text{ then}$$

$$F_1 + 330.75 - 294 = 0$$

$$\Rightarrow F_1 = -36.75 \text{ N, or } 37 \text{ N downwards}$$

Some interesting aspects of problem-solving arise from Worked example 7.8C. It was assumed that the forces supplied by both the pillars would be upwards. This assumption proved to be wrong, but because care was taken with the directions of the known forces, the correct directions for the forces emerged. Clearly, the beam needs to be attached to pillar 1, using nails, screws or bolts to hold it down. It is worth noting also that if the cantilever had been moved so that its centre of gravity lay between the two supports, the direction of 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 might 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.

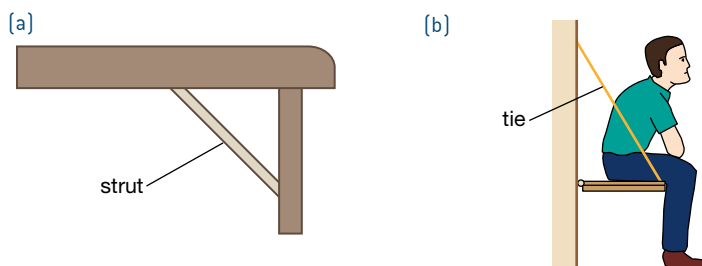


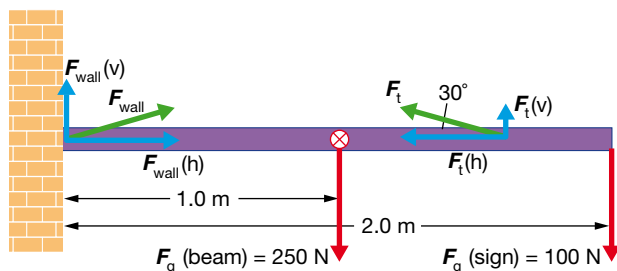
Figure 7.43 (a) A strut helps to support a cantilevered beam and is under compression. (b) A tie helps to support a fold-out bench and is under tension.

Worked example 7.8D

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 $g = 10 \text{ N kg}^{-1}$ 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 7.8C. The forces are not simply perpendicular to the beam, so components must be used to determine the forces. Begin by identifying all the forces acting on the beam in a force diagram. There are four forces to consider. The weight of the beam, $F_g(\text{beam})$, will act at the centre of gravity, 1.0 m from the wall. The tension from the wire F_t acts along the wire. The weight of the sign, $F_g(\text{sign})$, pulls downwards at the very end of the beam. The force that the wall exerts on the beam, F_{wall} , is not so obvious. Let us think about what would happen to the beam if this force was not present, i.e. if the wall collapsed or vanished at the point of contact. Were this to happen, the beam would fall down and swing to the left under the influence of the remaining forces. This indicates that the wall exerts a force that is acting to the right and upwards. These forces and their components are shown in the following diagram.

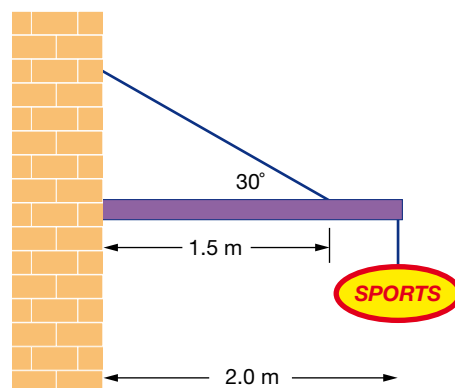


If the point at which the beam meets the wall is considered to be the pivot, then the weight of the beam and the sign supply clockwise torques, and the tension must supply an anticlockwise torque through its vertical component, $F_t(v)$. As the beam is in equilibrium, $\Sigma\tau = 0$ and $\tau_{\text{clockwise}} = \tau_{\text{anticlockwise}}$, so

$$(1.0 \times 250) + (2.0 \times 100) = 1.5 \times F_t(v)$$

$$F_t(v) = 300 \text{ N upwards}$$

$$\text{If the vertical component of the force is 300 N, } F_t = 300 \div \sin 30^\circ = 600 \text{ N}$$

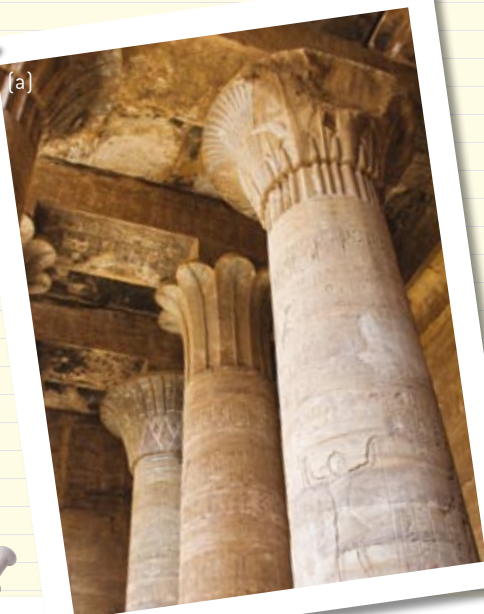


Structures through the ages

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: (i) it is very heavy, and (ii) it is weak under tension, even though it is very strong under compression.

(a)



(b)



Figure 7.44 (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 BC), 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 BC 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 hand-shaped stones 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 7.45 The aqueduct at Pont du Gard in the south of France. Built by the Romans in about 18 AD, 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.

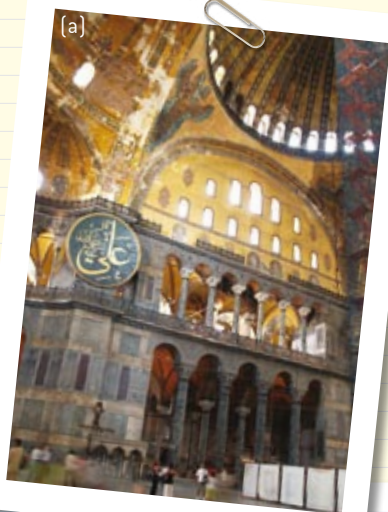


Figure 7.46 (a) The interior of the Hagia Sophia. Four piers in a square support four arches and the dome, which is clearly visible, rests on the arches. (b) On the exterior of the Hagia Sophia are a series of semi-circular domes that support the eastern and western walls of the building. A massive rectangular buttress (out of sight) is used on the northern and southern walls.

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 AD, 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 semi-circular arches rested. Resting on the arches was a large 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 halfdomes, 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 7.46b).

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.



Figure 7.47 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.

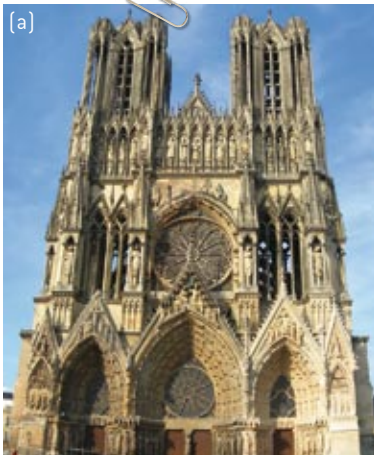


Figure 7.48 (a) Rheims Cathedral in France, whose construction began in 1210. (b) The pointed Gothic arch is a feature of many of the cathedrals built in Europe around this time. The horizontal forces created by the stone in the arch are not balanced by a wall, but are transferred to the outside of the 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.

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.

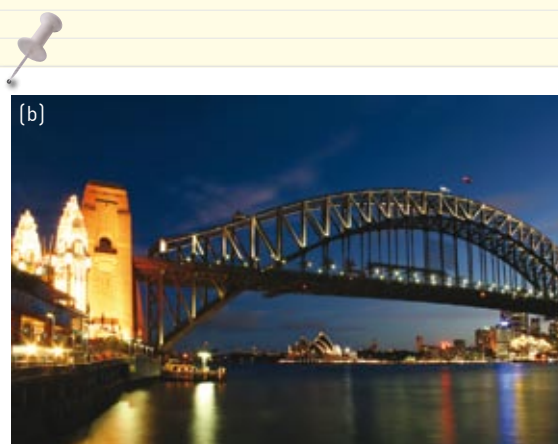
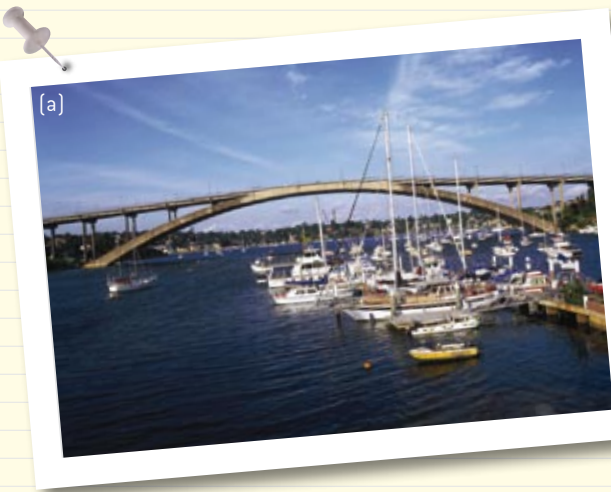


Figure 7.49 (a) The Gladesville Bridge across the Parramatta River. (b) The Sydney Harbour Bridge. Both bridges use an arch, but the designs produce different effects.



7.8 summary

Structures in translational and rotational equilibrium

- 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. $\Sigma F = 0$ and $\Sigma\tau = 0$. As a consequence of the translational equilibrium condition, $\Sigma F = 0$, $\Sigma F_x = 0$ and $\Sigma F_y = 0$ must also be true.

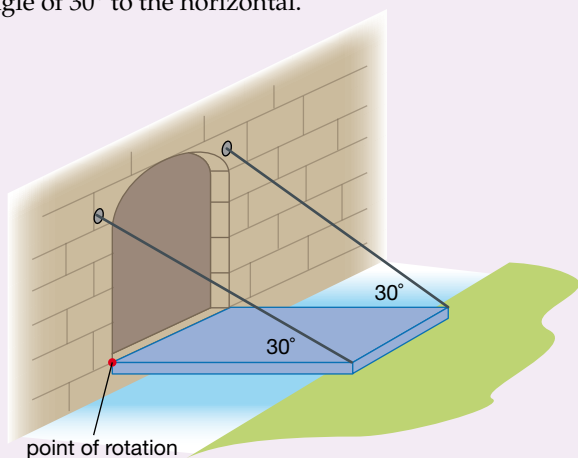


7.8 questions

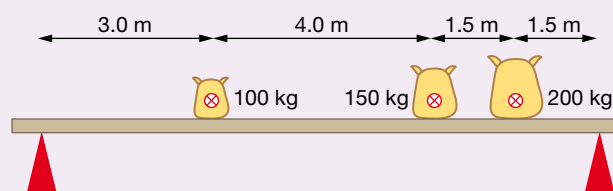
Structures in translational and rotational equilibrium

Use $g = 9.80 \text{ N kg}^{-1}$ when answering these questions.

- 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 mass of the second child?
- Explain how an adult can use a playground see-saw with a child so that the end that the child sits on is not always raised in the air, but remains horizontal.
- 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.
- 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.



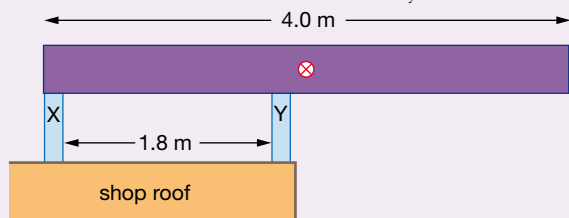
- Write an expression for the torque produced by each of the tensile forces that act around the axis of rotation of the bridge.
 - Write an expression for the horizontal and vertical components of the tension force.
 - Calculate the size of the tension in each cable.
- A makeshift shelf in a farm shed is constructed using a 10 m beam of mass 50 kg supported at each end. If the shelf supports three sacks of wheat of mass 100 kg, 150 kg and 200 kg at the positions shown, calculate the support forces at each end of the beam.



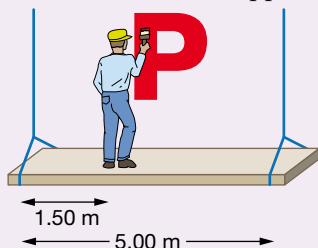
- A train engine passes over a 20 m bridge span which is supported by two columns X and Y. The engine has a total mass of 5.0 tonnes. At one instant column Y produces a reaction force of 30.6 kN. If the spanning beam is uniform and has a mass of 5.0 tonnes, where is the centre of mass of the train?
- 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:
 - the weight of the ladder
 - the weight of a person of mass 50 kg standing one-quarter of the way up the ladder
 - the weight of a person of mass 50 kg standing three-quarters of the way up the ladder.



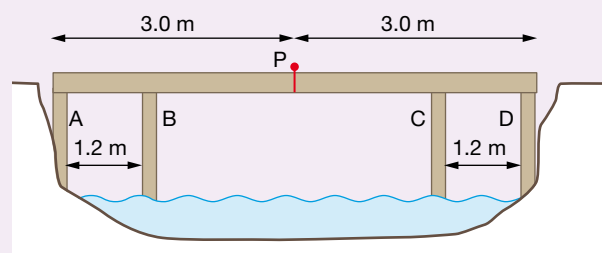
- 8 A 4.0 m cantilever-type verandah is constructed on the roof of a shop. The verandah has a mass of 900 kg and is supported by two supporting columns X and Y, which produce reaction forces F_x and F_y , respectively.



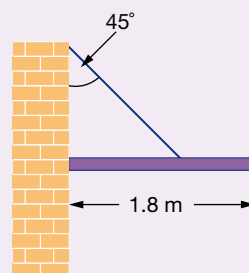
- Write an equation which shows how all vertical forces are balanced.
 - Write an expression relating the torques about the point where the beam contacts pillar X.
 - Determine the force supplied by column Y, F_y .
 - Write an expression relating the torques about the point where the beam contacts pillar Y.
 - Determine the force F_x and indicate whether X and Y are under compression or tension.
- 9 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.



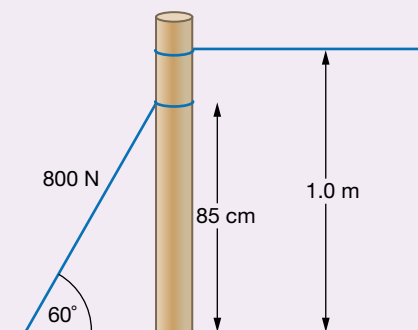
- The painter alters his position so that the left-hand rope now experiences a tension of 557 N and the other a tension of 325 N. Where is the painter now standing in relation to the left-hand rope?
- 10 A pedestrian bridge over a small creek is made from two identical 3.0 m long cantilevers, each of mass 400 kg.
- Calculate the reaction forces produced by the pillars A, B, C and D
 - when there are no pedestrians on the bridge
 - when a 70 kg person stands at position P with half her weight on each cantilever.



- What happens to the values of the forces in A and B as the woman walks from A past B to P?
- 11 A uniform 5.0 kg beam, 1.8 m long, extends from the side of a building and is supported by a cable which is attached 1.2 m from the wall at an angle of 45° . Determine the tension in the cable.



- 12 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.



- Determine values for the horizontal and vertical components of the tension in the backstay wire.
- By considering the base of the post to be a pivot point, determine the size of the tension in the fence wire, F_t .



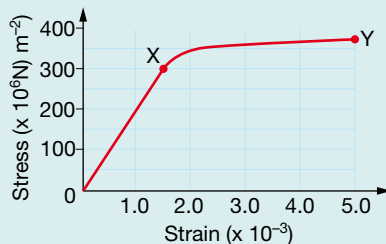
chapter review

Multiple-choice questions

Use $g = 9.80 \text{ N kg}^{-1}$ when answering these questions.

The following information applies to questions 1–6.

The graph shows stress versus strain data for a section of metal wire of cross-sectional area $2.0 \times 10^{-6} \text{ m}^2$ and initial length 1.50 m. At X the wire lost its elastic properties, and at Y the wire broke.



- Which option best gives the tensile strength of the metal used in this wire?
A $3.5 \times 10^8 \text{ MPa}$ **B** 1.5 MPa **C** 1.8 MPa
D 350 MPa **E** 300 MPa
- What is the elastic limit of this metal?
A 350 N m^{-2} **B** $3.5 \times 10^8 \text{ N m}^{-2}$ **C** $3.0 \times 10^8 \text{ N m}^{-2}$
D 300 N m^{-2} **E** 5.0 N m^{-2}
- Which is the value of Young's modulus for this metal?
A 370 N **B** 300 N **C** $2.0 \times 10^{11} \text{ N m}^{-1}$
D $1.5 \times 10^8 \text{ Pa}$ **E** $1.5 \times 10^{10} \text{ Pa}$
- Which option best represents the amount of energy this section of wire absorbed while being stretched to its elastic limit?
A 225 kJ **B** 1.4 kJ **C** 14 J
D 0.68 J **E** 300 J
- Which option best represents the amount of energy this section of wire absorbed before it broke?
A 1.4 MJ **B** 1.4 kJ **C** 14 J
D 2.3 J **E** 4.2 J
- The behaviour of the wire when it is stretched past the elastic limit is best described as:
A elastic **B** plastic **C** stiff
D brittle **E** tactile.

The following information applies to questions 7 and 8.

A 6.0 m ladder of mass 7.5 kg leans against a wall at a 65° angle to the horizontal.

- 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 200 N m **E** 94 N m
- A 60 kg woman is standing one-third of the way up the ladder. What is the torque (taken around where the ladder rests on the ground) exerted on the ladder by the woman?
A 500 N m **B** 22 N m **C** 31 N m
D 200 N m **E** 93 N m

The following information applies to questions 9 and 10.

A crane is lifting a prefabricated concrete wall of mass 4.5 tonnes. Assume that a single steel cable is being used to lift the load, and ignore the mass of the cable in your calculations.

- 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
- The load is now lifted at a constant speed of 2.0 m s^{-1} . Which of the following best describes how the tension acting in the cable compares with the tension value in Question 9?
A The tension in the cable is now greater.
B The tension in the cable is equal to that in Question 9.
C The tension in the cable is less than that in Question 9.
D This cannot be determined from the information given.

Extended-answer questions

- A 20 cm length of brass wire of radius 1.0 mm experiences an elongation of 0.49 mm when a tensile force of 1.0 kN is applied.
a Calculate the value of Young's modulus for brass.
b Determine the percentage strain of the brass.
c Explain why Young's modulus is a more useful quantity than the spring constant when describing the properties of a material.

The following data for concrete applies to Question 12.

Tensile strength = $2.0 \times 10^6 \text{ N m}^{-2}$

Compressive strength = $2.0 \times 10^7 \text{ N m}^{-2}$

Young's modulus (compression) = $2.0 \times 10^{10} \text{ N m}^{-2}$

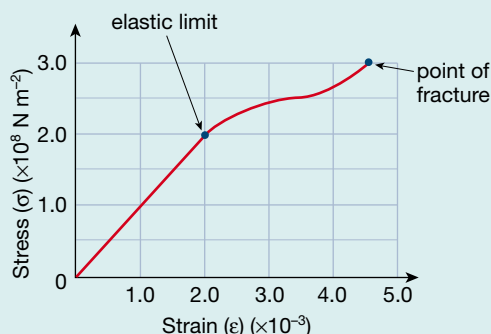
Elastic limit = $2.0 \times 10^6 \text{ N m}^{-2}$

- a** Calculate the minimum radius of a cylindrical concrete column that could just support a load of 10 tonnes without damage.
b What is the maximum amount of energy that a concrete wall of volume 4.0 m^3 could absorb under compression before exceeding its elastic limit?
c Which statement best explains why concrete by itself is a more suitable material for use in columns than beams?
A Concrete is stronger under tension than under compression.

- B** Concrete is stronger under tension than under shear stress.
- C** Concrete is stronger under shear stress than under compression.
- D** Concrete is stronger under compression than under tensile stress.
- d** What modification is used to enable concrete to be used safely in beams and floors?

The following information applies to questions 13–15.

A new alloy is being tested for use in space. The stress–strain graph for this material under compression is shown.



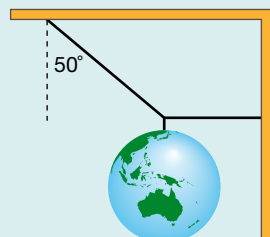
- 13** What is the compressive strength of this alloy?
- 14** A cylindrical rod of length 8.00 cm and cross-sectional area 1.0 cm^2 is made from this alloy.
- Calculate the compression in the rod under a stress of $2.0 \times 10^8 \text{ N m}^{-2}$.
 - What is the maximum compressive force that can be applied to the rod before it fails?
 - What is the maximum compressive force that should be applied to the rod if a safety factor of 5 is used?
- 15** Calculate the amount of strain energy the rod in Question 14 can absorb without failing.
- 16** Young's modulus and the tensile strength for three different materials are as follows.

Material	Young's modulus (N m^{-2})	Tensile strength (N m^{-2})
Steel	2.0×10^{11}	8.2×10^8
Aluminium	7.0×10^{10}	2.0×10^8
Nylon	7.0×10^8	5.0×10^8

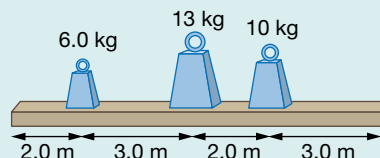
- Explain the term 'stiffness' with respect to the physical properties of a material.
- Rank these materials in order of increasing stiffness.
- A piece of each of these materials is subjected to a stress of $1.0 \times 10^5 \text{ N m}^{-2}$, a value well below the elastic limit for each material. Calculate the strain energy per unit volume in each sample.
- From the data, it might be fair to say that steel is a 'tougher' material than aluminium. Assuming this to be true, what

conclusion could you make concerning a comparison of the stress–strain graphs of these materials?

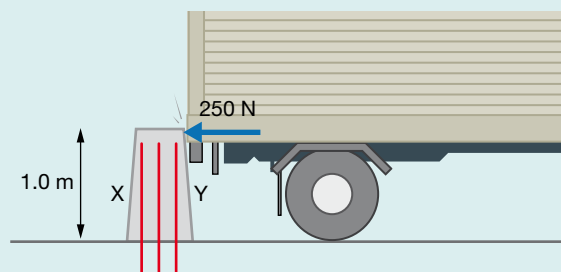
- 17** A 1.5 kg model of the Earth is suspended by two long wires in a school library as shown. Calculate the tension in each wire.



- 18** A 2.0 kg beam supports masses of 6.0, 10 and 13 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.



- 19** A crane with a horizontal arm is lifting a steel girder of mass 1.5 tonnes. Initially, the load was being lifted at the far end of the horizontal arm, a distance of 25 m from the pivot of the crane. Ignore the mass of the cable when answering these questions.
- The driver decided that it would be wise to bring the load in closer to the pivot. The crane carries a counterweight of mass 20 tonnes. How far from the pivot should this counterweight be positioned so that its torque balances the torque of the load?
 - What is the benefit of bringing the load in closer to the crane pivot?
- 20** A barrier 1.0 m high in an underground car-park is a composite material made of concrete and steel reinforcing rods. A truck reverses into the barrier and exerts a force of 250 N on it, as shown in the diagram.



- Calculate the magnitude of the torque around the base of the barrier that this force creates.
- Is the concrete more likely to crack at X or Y as a result of this collision? Explain.

Further electronics

To say that electronic systems are all around us is almost an understatement. A little appreciated fact, however, is that while virtually all of our electricity is produced as high-voltage alternating current, these electronic systems are dependent on a low-voltage direct current supply. The conversion of the former to the latter, with minimal loss of energy, is therefore one of the most important tasks any electrical engineer has to accomplish. Many of the devices we seem to find so indispensable these days come with a small 'power supply' which performs this function—often charging batteries which provide the mobile low-voltage DC to our phones, computers, cameras and a host of other such devices. Other electronic devices such as music systems, desktop computers, TV sets and a huge array of sensors and control systems will have built-in power supplies. In this chapter we work from the ideas in Chapter 4 and extend them to the construction and use of a low-voltage DC-regulated power supply.

One of the purposes of this detailed study is to give you a chance to construct a circuit yourself. There are many possible ways to go about the construction. There are kits complete with instructions and all the parts mounted on boards that clip together. Or you can buy all the individual components from your local electronics store and solder them together on nails in a block of wood. Whatever method you use, the most important thing is to understand the principles behind the circuit. It is an understanding of these principles that will enable you to improve and adapt circuits, or to design completely new circuits for different purposes.

by the end of this chapter

you will have covered material from the study of electronics, including:

- the design of a transformer-supplied AC-to-DC voltage-regulated power supply system
- the effects on the system, as interpreted from a CRO display, of changes to the components used
- the use of measuring devices (including a CRO) to analyse the system and to diagnose faults
- the operation of diodes in half-wave and full-wave rectification
- the effects of capacitors in circuits in terms of their charging and discharging time constant and in smoothing DC
- the use of zener diodes and voltage regulators in power supply circuits
- the effects of a change in the load on the ripple voltage
- power dissipation and the use of heat sinks in electronic circuits.

outcome

On completion of this chapter, you should be able to design and investigate an AC-to-DC voltage-regulated power supply system, and describe and explain the operation of the system and its components, and the effects of test equipment on the system.

CHAPTER 8

The basic principles of designing an electronic device are common to all electronic circuits. Whether it is a low-voltage power supply, a stereo amplifier, or a complex digital circuit, many of the basic components are the same and they work in similar ways. Let's consider, for a moment, the construction of a *control device*; that is, a system which responds to a change in some quantity by turning something else on or off. Simple examples are a motion sensor light, a thermostat, a burglar alarm, a washing machine water-filler system, a street-light controller and a garden watering system. All of these devices are effectively made up of four parts:

- a sensor that provides an electrical signal input
- a system that converts information from the sensor into an appropriate electrical signal
- an output that performs the required task on receiving the electrical signal
- a power supply.

A stereo system amplifier is not dissimilar. The low-level input signal is provided from a CD player, a radio tuner or other source. The purpose of the amplifier is to take this signal and amplify it so that the output signal is:

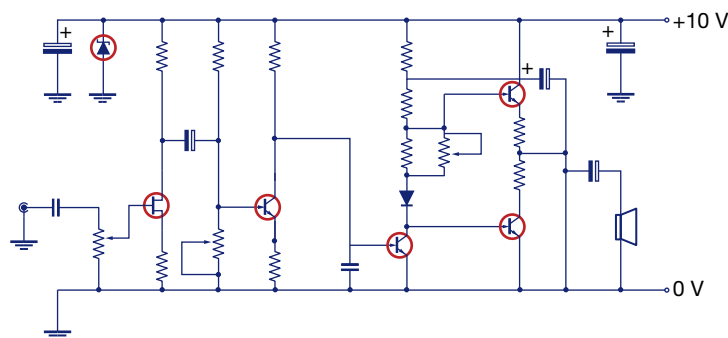
- a faithful reproduction of the original low-level signal
- powerful enough to drive the loudspeakers.

One of the most important components of the amplifier is the power supply. It must feed the rest of the amplifier with a constant DC voltage which doesn't alter, whatever the output—from the quiet whisper of a peaceful song to the crashing sound of a rock band or symphony orchestra. Any variation in the voltage supplied will result in distortion of the sound.



Figure 8.1 Virtually all electronic devices rely on a power supply to provide a steady low-voltage direct current.

Figure 8.2 In circuit diagrams such as this one, the power supply is represented simply by the +10 V and 0 V lines, a common way to represent the power supply in electronic circuit diagrams.



In an electronic circuit diagram such as that in Figure 8.2, the power supply is shown by the +10 V line at the top and the 0 V line at the bottom. This is a common way to set out such diagrams. The power supply in this diagram could just be a battery, it could be a solar cell, or it could be, and often is, a circuit that takes the mains 240 V AC and converts it to a low-voltage DC. Even in many devices operated by a battery, the battery will be recharged by such a power supply. The aim of this study is to investigate the principles of, and then construct, a power supply of this type. The power supply is really a sort of 'unsung hero' of modern electronics. It must supply a constant and smooth DC voltage despite variations in both the mains supply and the current drawn by the circuit. In many cases it must operate 24 hours a day, 7 days a week, and draw hardly any current while the system is left on 'standby'—as so many modern devices are. In this and the next section of this chapter we will look at some of the components and principles involved, and then in section 8.3 we will look at the design and construction of a typical power supply.

Basic principles

It is important to recall the basic principles behind the analysis of any electrical circuit. These were studied in Year 11 and again in the Area of study 'Electronics and photonics'.

Kirchoff's two laws express the fundamental laws of conservation of energy and conservation of charge, and are the basic principles used in the analysis of any electrical circuit:

- 1 In any electrical circuit the sum of all currents flowing into any point in the system is zero.
- 2 The total potential drop around any closed path in a circuit must be equal to the total EMF in the path.

The first law tells us that current does not get 'lost', or build up, at any point in the circuit. At any intersection of conductors, the total current flowing in is equal to the total current flowing out (hence zero total current). The second law tells us that as we move around through the circuit from one terminal of the power supply to the other, the total of the drops in potential across the various components in the circuit must be equal to the voltage of the power supply. In most circuits of the type we will be discussing, the positive power supply line is drawn along the top of the diagram and the negative—which is often connected to 'ground'—along the bottom. In practice, this means that the total voltage drop down any line from top to bottom must equal the EMF of the power supply.

AC and DC

DC power supplies, such as batteries, cause the free electrons in a circuit to drift constantly in only one direction. Large electric power generators, such as those in the Latrobe Valley, always produce AC, or alternating current.

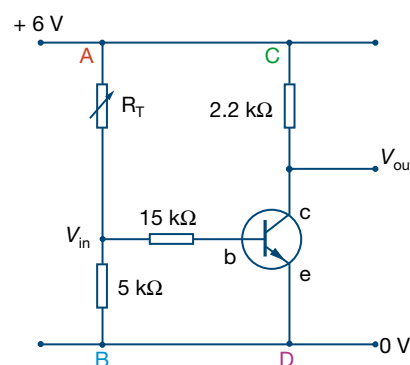


Figure 8.3 The total voltage drop down any path from top to bottom in this simple transistor circuit is equal to the supply voltage. For example, the total voltage drops from A to B through R_T and the 5 k Ω resistor are equal to the total drops from C to B through the transistor and the 2.2, 15 and 5 k Ω resistors, and both are equal to 6 V.

Figure 8.4 In a conductor carrying DC, the electrons move steadily in one direction. In AC, they oscillate back and forth.

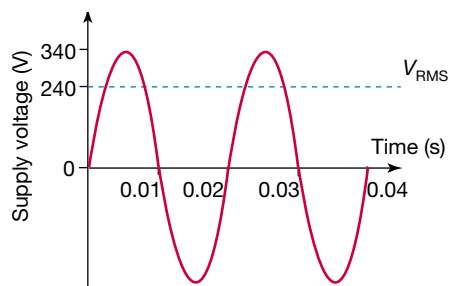


Figure 8.5 '240 V' in relation to the normal domestic supply refers to the root mean square [RMS] value. The corresponding peak value is about 340 V.

Physics file

Peak values and RMS values are often quoted for sound systems. The peak power output refers to the maximum power output produced during a cycle by the particular amplifier and loudspeaker system during testing by the manufacturer. This is the power produced only at the peak of an oscillating voltage cycle, so it is actually quite misleading. The RMS value, which in simple terms is the effective average maximum output of the system (and is generally half the peak power), is a more useful specification when comparing the output of different sound systems.

The current in an AC circuit changes direction many times each second. The free electrons in the conducting wires move first in one direction and then the other, creating an oscillating voltage 'wave'. The term AC refers to the current, but of course the reason the current oscillates is that the voltage driving it is oscillating.

Figure 8.5 shows how the voltage of normal 240 V AC mains supply changes with time. The peak value actually swings between about plus and minus 340 V. This is sometimes referred to as a peak-to-peak voltage of 680 V. Describing the voltage and current values for an AC circuit is not straightforward because they are always changing. The average value cannot be used because it is generally zero for both voltage and current, and the maximum value is not appropriate because it is reached at only two instants in each cycle. The most useful measurement of an AC voltage is what is called its *root mean square* (RMS) value. This term refers to the mathematical relationship involved, the details of which can be found in section 10.5. Importantly for us, the RMS voltage is equivalent to the DC voltage that would be needed to produce the same power. For example, by saying that an AC circuit has an RMS voltage of 240 V we mean that it will produce the same power as a 240 V DC supply in a simple circuit consisting of circuit elements which act only as resistors. Voltage or current values quoted for AC circuits can generally be assumed to be RMS values (sometimes also called *effective values*) rather than peak values.



In an AC circuit the **RMS VOLTAGE** is defined as $V_{\text{RMS}} = \frac{V_{\text{peak}}}{\sqrt{2}}$
and the **RMS CURRENT** as $I_{\text{RMS}} = \frac{I_{\text{peak}}}{\sqrt{2}}$.

The period of an AC voltage is determined by the rate at which the generators turn. In Australia, and in many other countries, this rate is carefully controlled by the power station operators so that the frequency of the mains voltage is exactly 50 cycles per second, or 50 Hz. Hence, the period is 0.02 s, or 20 ms. (So in 1 second, the electrons in a conductor connected to the AC mains would change direction 100 times!)

Provided an AC voltage is used in a circuit containing normal resistances (light bulbs, heating elements, metal film resistors and so on) Ohm's law can be used to calculate the current that will flow. However, if the circuit contains coils that produce significant magnetic effects, Ohm's law cannot be used. Electric motors, loudspeakers, transformers and relays are some examples of components that produce such magnetic effects. (We will discover the reason for this when we study electric power in Unit 4.) It is important to keep this in mind because if a DC voltage is used instead of a similar AC voltage, the current through such devices will be far higher and can burn them out.

In order to power a light bulb or a heater, either AC or DC can be used. The heat and light generated are not dependent on which way the current is flowing. Some electric motors are designed to work on either AC or DC, but most are designed to work only on AC or only on DC. However, the circuits in amplifiers, computers and indeed virtually all electronic devices, must be supplied with a smooth, constant, low-voltage DC power. It is therefore important to be able to convert the AC mains power into low-voltage DC power. We will discover how this can be done later in this study.

For more details on electric power see section 10.2.



Figure 8.6 It is important to use the right type of electric power to run any electrical device.

Practical electronics

In any practical electronics project, it is important to be familiar with the basic tools used to make measurements and to 'see' the voltages and currents in a circuit. Any test instrument will affect the circuit being tested to some extent. A voltmeter (or multimeter being used as such) will take a little current from the circuit in order to either move the coil to which the needle is attached, or operate the digital circuits. Digital meters normally take much less current, so when you are measuring voltages in circuits where the currents flowing are very small, it is best to use a digital meter if possible.

Likewise, an ammeter will have a very small resistance to the current flowing through it, and therefore a small potential drop across its terminals. This may be significant in situations where relatively low voltages are being measured. As the power supply circuit we shall be building has relatively high currents, normal (moving coil) multimeters or separate volt and amp meters will be adequate.

Increasingly, the CRO is being superseded by the use of computer-based equivalents. These work in a similar way, but have the advantage of features that enable them to capture a waveform. A simple CRO will only show us the voltage waveform in a circuit as it happens, but sometimes the waveform may be too quick to see. A digital or computer-based CRO is able to store a rapidly changing voltage pattern and then display it later so that we can examine it at our leisure.

Electronic components

The other essential information we require before we start on any electronic project is an understanding of the basic components we will be using. In Chapter 4 we studied most of the important components that are used in electronic circuits. Here we will list them again and describe some of their more important features as used in the practical circuits we will be concerned with.

Physics file

It is important to be familiar with some basic techniques for using multimeters in a circuit. Remember that ammeters and voltmeters are used very differently—the voltmeter across (i.e. in parallel with) a circuit element and the ammeter in series with it. A multimeter can act as either a voltmeter or an ammeter. For this reason, always disconnect the meter from the circuit before changing the function. You may also need to change the lead connections on the meter when changing functions. Make sure that you are familiar with the multimeter's functions before using it. When using an ammeter, there is a danger of short-circuiting something as it is, effectively, an almost zero resistance connection. Be very careful to avoid connecting an ammeter into a circuit as though it was a voltmeter! In this situation you may damage the circuit, the multimeter or both. An ammeter must be in *series* with part of the circuit.

When using the multimeter as an ohmmeter, remember that the meter puts a small current through the object being measured. It is important to disconnect at least one end of the resistor being measured from the circuit before making a measurement or you will be measuring the resistance of a whole circuit.

While computer-based measurements may look different to traditional meters, the basic principles outlined above still apply.

For more information on cathode ray oscilloscopes and multimeters, see Chapter 4.



PRACTICAL ACTIVITY 27

Using electrical meters

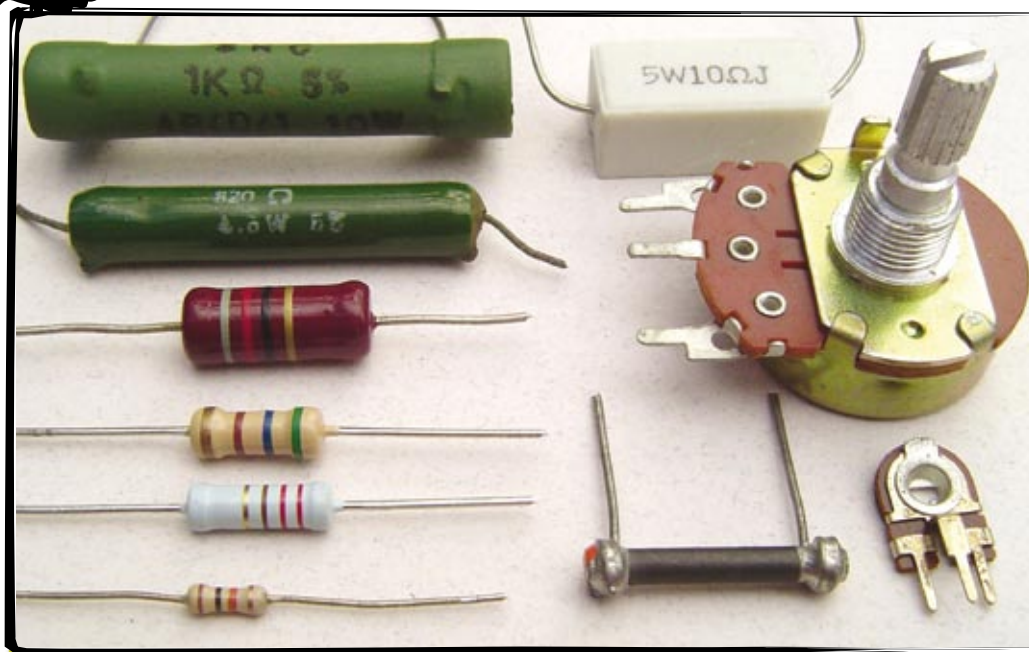


Figure 8.7 Many resistors have coloured bands to represent their resistance value. Others have the value printed on them. Also shown are a thermistor (lower middle) and two potentiometers or variable resistors.

Resistors

Electronic circuits use either carbon-film or metal-film resistors to carry out functions such as controlling the voltage across a particular part of a circuit. Carbon-film resistors, which are available in a large range of resistance values, are sometimes called ‘carbon-composite resistors’ since they are made of powdered carbon mixed with a glue-like binder. Metal-film resistors are made by depositing a thin film of metal onto a ceramic core and then applying a protective coating. Most resistors are colour-coded so that the resistance value can be easily identified. (See the Physics in action on page 287.)

The physical size of these resistors is not related to their resistance value. Rather, the size of a resistor relates to the electrical power that it can handle. Low-resistance resistors are often physically larger than high-resistance ones, as they are designed to handle greater current and therefore higher power loads. Resistors convert electrical energy to heat, and if a resistor gets too hot, it may not operate correctly or it may even burn out completely. It is, therefore, important to take note of the power ratings of resistors. Most of the small resistors used in circuits are $\frac{1}{4}$ watt, but $\frac{1}{2}$ watt and 1 watt are also common. Remember that the power dissipated in any device depends both on the current through it and the voltage across it ($P = VI$).

Wire-wound resistors—as their name implies—are made by winding a thin wire around a core and coating it in a heat-conductive ceramic material. They are generally used where higher power dissipation is needed or where small values of resistance are required.

Resistors are normally ohmic devices, at least if operated within their specified power ratings. However, not all resistive devices have a constant resistance. The resistance of a simple light bulb, for example, increases considerably when it is on. Components known as *thermistors* have the opposite characteristic. The hotter they become, the lower their resistance. They are commonly used as temperature sensors. The resistance of light-dependent resistors (LDRs) decreases with the amount of light. This enables them to be used to detect or measure light.

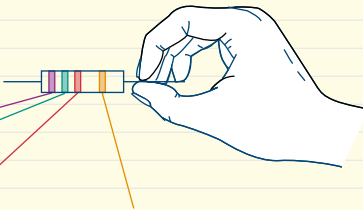
The international resistor colour code

Because resistors are normally too small to have the value of their resistance and tolerance printed on them, an international colour code was developed, using four coloured bands around the resistor. The colour of each of these bands represents a number which can be used in determining the value of the resistance and its manufactured tolerance. Bands 1 and 2 represent the two digits of a number in the range from 00 to 99. Band 3 is the multiplier—the number of zeros appearing after the two digits. Band four is the resistor's tolerance.

It is difficult to produce a carbon resistor with an exact value. Its resistance value can only be quoted to within a

certain range. This is the tolerance of the resistor, typically 5% or 10%. For example, a $22\ \Omega \pm 10\%$ resistor could have a resistance anywhere between $20\ \Omega$ and $24\ \Omega$. This is one reason why manufacturers don't bother producing 20, 21, 23 or $24\ \Omega$ resistors. Consider also the economics of producing a resistor for each value between $0.1\ \Omega$ and $10\ \text{M}\Omega$!

Resistor values have been determined so that a complete range (from $0.1\ \Omega$ to $10\ \text{M}\Omega$) is covered by 97 different resistors. Typical values are 10, 12, 15, 18, 22, 27, 33, 39, 47, 56, 68 and $82\ \Omega$, and decimal multiples of these (e.g. $8.2\ \Omega$, $220\ \Omega$, $27\ \text{k}\Omega$, $470\ \text{M}\Omega$ and so on).



1st colour band 1st digit	2nd colour band 2nd digit	3rd colour band Number of zeros	4th colour band Tolerance
Black 0	Black 0	Black 0	Gold 5%
Brown 1	Brown 1	Brown 1	Silver 10%
Red 2	Red 2	Red 2	
Orange 3	Orange 3	Orange 3	
Yellow 4	Yellow 4	Yellow 4	
Green 5	Green 5	Green 5	
Blue 6	Blue 6	Blue 6	
Violet 7	Violet 7	Violet 7	
Grey 8	Grey 8	Grey 8	
White 9	White 9	White 9	

Figure 8.8 The international resistor colour code.

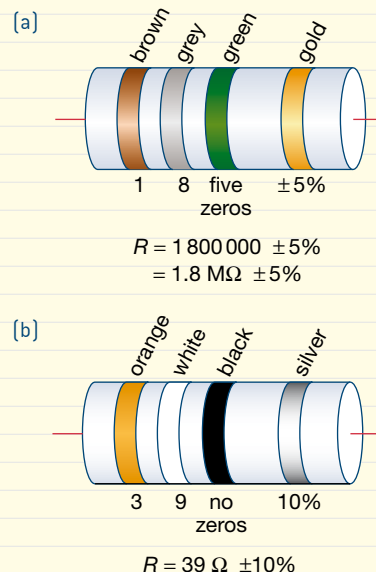


Figure 8.9 Examples of the resistor colour code: (a) a $1.8\ \text{M}\Omega \pm 5\%$ resistor, and (b) a $39\ \Omega \pm 10\%$ resistor.

Diodes

The key feature of a diode is that it allows current to flow in one direction but not the other. This property will be crucial in our construction of a power supply which takes AC current flowing alternately in both directions, and produces DC current flowing in only one direction. Figure 8.10 shows the general I - V characteristics of a diode. Diodes can be characterised by specific properties which are of importance in particular applications. Three of these are as follows.

- V_s is the voltage at which the diode starts to conduct significantly. It is important to remember that this voltage drop, or more, exists across a diode even when it is conducting well.
- When reverse biased, a very small 'leakage' current does flow, although for most purposes this will be negligible—typically only a fraction of a microamp.
- At some reverse voltage (hopefully fairly high) the diode will break down (V_b) and a high reverse current will flow. This usually results in the destruction of the diode as the power released will be considerable.

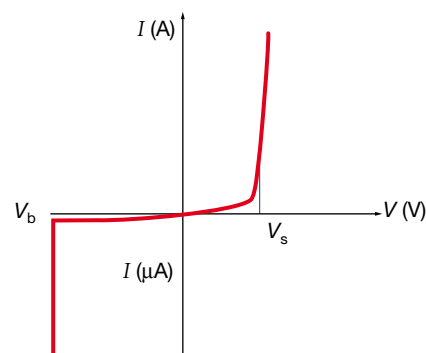


Figure 8.10 The general characteristics of a diode. V_s is the voltage at which it starts to conduct significantly, V_b is the breakdown voltage, the reverse voltage at which the diode will conduct in the reverse direction. The reverse [leakage] current is highly exaggerated on this graph. If it was drawn to the same scale as the forward current it would not be visible.

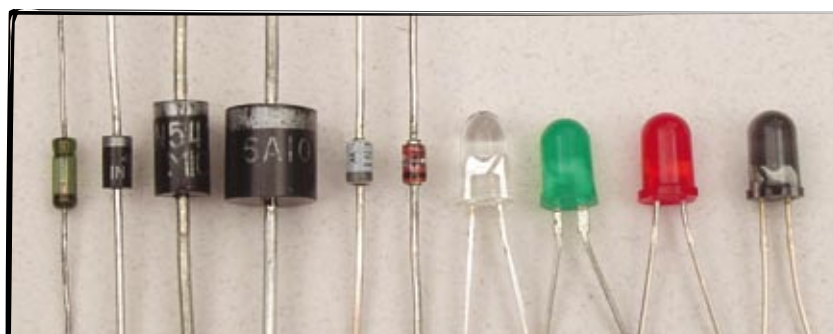


Figure 8.11 Diodes come in many forms. Here we see a small signal diode, three power diodes, two zener diodes and four LEDs.

There are many types of diodes made for different purposes. We will look briefly at two types which will be useful to us in this study. *Power diodes* are made for use in AC to DC power supplies. Their function is to allow the current to flow one way only. The primary requirement of a power diode is that it can handle large forward currents with as little voltage drop (V_s) as possible. Remember that because of this voltage drop, power ($P = V_s I$) will be produced in the diode. This will limit the amount of current that can be provided. Power diodes are therefore made in a very large range of current-handling capacities. Generally, the larger the diode, the better able it is to disperse the heat generated, and so the larger its current rating. Many power diodes are made either with an integrated heat sink or in such a way that they can be mounted on one. A *heat sink* is usually a black piece of aluminium that conducts heat away from the diode.

Zener diodes have relatively low reverse-breakdown voltages. This may seem to be an undesirable characteristic, but in fact they are made with a very precise breakdown voltage that can be used to regulate the voltage in a circuit. When used in this way, they are rather like a safety valve which releases excess pressure. We shall look at their use for this purpose in section 8.3. Used in the correct way, the reverse current will be limited by the circuit and the zener diode will not be destroyed. Zener diodes are manufactured with a range of zener voltages and power ratings.

Capacitors

Capacitors are a basic component of any power supply. If one looks at the circuit boards inside an old radio or TV set (never look inside a functioning one!) we see many little green, blue or brown, two-terminal rectangles, disks or cylinders. Many of these will be capacitors. Clearly they have a very important role in modern electronics. There is a simple reason for this. All of these devices deal with 'signals'—electrical voltages or currents that vary with time. Any signal that conveys information must change with time. Whether it is the voltage produced by a microphone that represents human speech, the rapidly varying signal that paints the picture on the screen of a TV, or the ultrafast digital data flying around inside a computer, a signal is basically a voltage (or current) that varies with time. In order to handle this sort of signal, it is often necessary to separate the varying voltage (AC) from a steady (DC) voltage. Capacitors can achieve this. Sometimes it is necessary to do the reverse, to remove a varying signal. Again, capacitors will do the job. We will look at them in more detail in the next section.

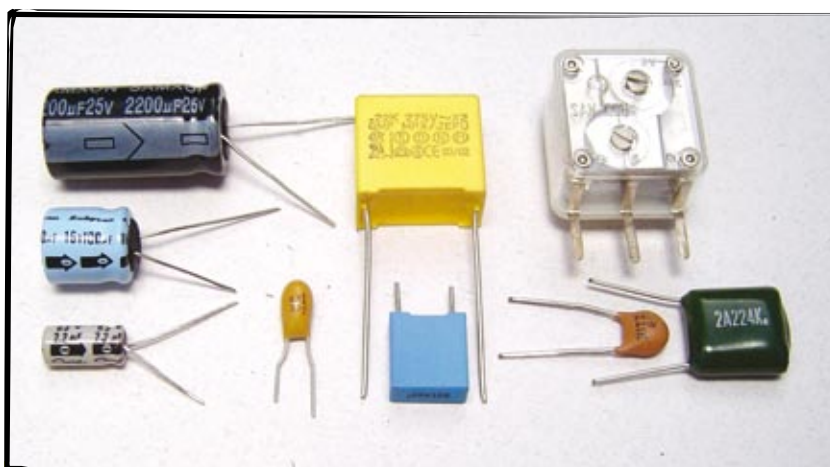


Figure 8.12 Capacitors come in many shapes and forms. Shown are some electrolytics, a tantalum, polyesters, ceramic and a variable (tuning) capacitor. Can you identify them?

Physics in action

Transistors

Although we will not be using individual transistors in this study, they are at the heart of most modern technology. From the tiny units that are etched into an integrated circuit in their millions, to large power-handling transistors, they control, detect, amplify and switch currents ranging from nanoamps to hundreds of amps. In fact, the voltage regulator that we will use later is actually an integrated circuit which incorporates about 16 transistors.

A typical transistor found in simple electronic circuits might be labelled something like 'BC548' or '2N3536'. It will come with a set of specifications which can be rather confusing to the beginner, but which are needed if the transistor is to be used correctly in a circuit. Table 8.1 shows a set of typical specifications for a transistor. We will look at the meaning of these various quantities.

Table 8.1 Specifications for a BC548 transistor

Transistor	BC548
Case	T0-92
Type	N
Material	Silicon
V_{ce}	30 V
V_{cb}	30 V
I_c	100 mA
V_{ces}	0.6 V @ I_c 100 mA
H_{fe}	110–800 @ I_c 2 mA
F_T	300 MHz @ I_c 10 mA
P_{tot}	500 mW

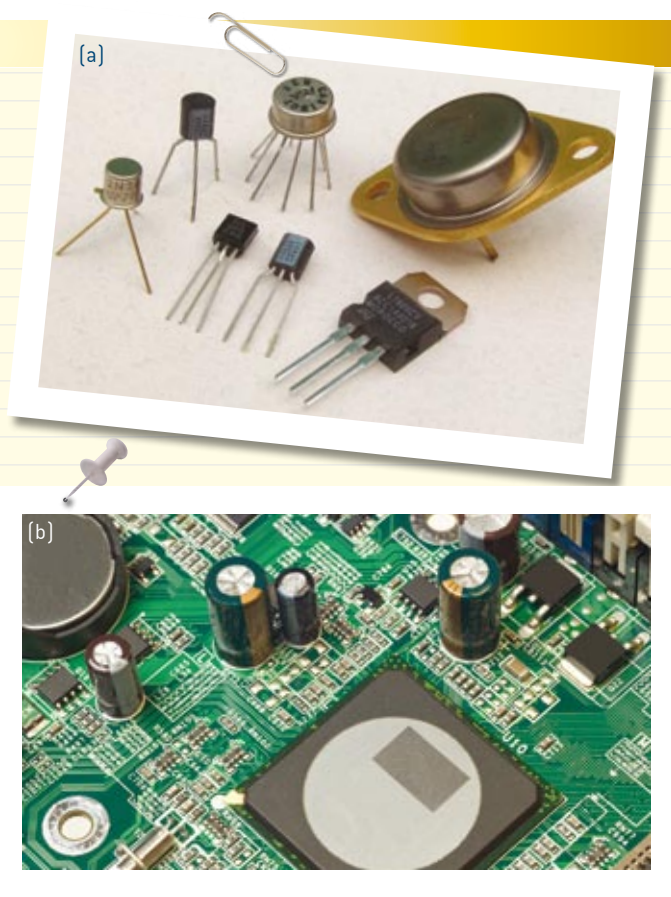


Figure 8.13 (a) Transistors come in many forms: individually in three-terminal packages, or packaged as integrated circuits. (b) The CPU of a modern computer contains millions of transistors.

The first three specifications tell us the type of case the transistor is built into, and the fact that it is an n-p-n silicon transistor. The V_{ce} and V_{cb} figures are the maximum voltages that can be put across the collector and emitter, and collector and base terminals, respectively. I_c is the maximum collector current that the transistor can handle. V_{ces} is the typical value of the voltage between the collector and emitter when the transistor is on, in this case with a current of 10 mA flowing. H_{fe} is the current gain of the transistor. This varies widely, but when used in a circuit, the overall circuit gain is set by the amount of negative feedback incorporated into the circuit. This involves choosing appropriate resistors which feed some of the output back to the input in such a way as to limit the amount of gain achieved and to make the circuit more stable. The associated I_c value indicates a maximum current at which the gain is achieved. F_T indicates the upper limit of frequency at which the transistor can be used effectively. P_{tot} is the total power that the transistor can handle and will basically be given by the collector current multiplied by the voltage across the collector and emitter.

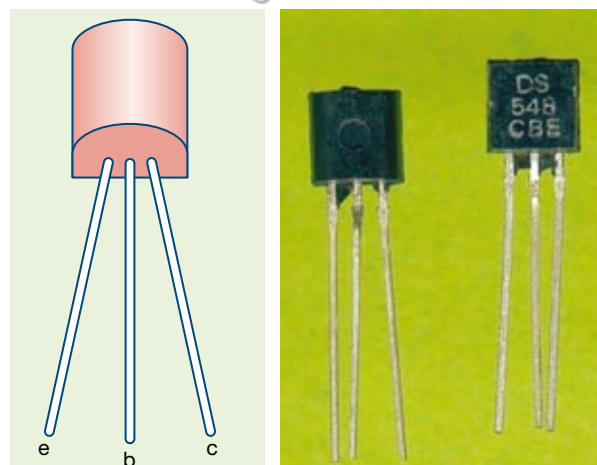


Figure 8.14 The pin connections of a transistor.



8.1 summary

Principles and practicalities of electronic design

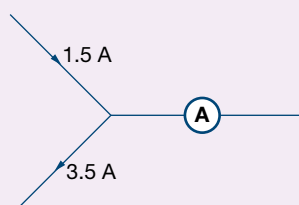
- Electronic control systems typically consist of an input, a system to manipulate the electrical signal, an output, and a power supply.
- Kirchoff's first law states that in any electrical circuit the sum of all currents flowing into any point in the system is zero.
- Kirchoff's second law states that the total potential drop around any closed path in a circuit must be equal to the total EMF in the path.
- In an AC circuit the RMS voltage is defined as $V_{RMS} = \frac{V_{peak}}{\sqrt{2}}$ and the RMS current as $I_{RMS} = \frac{I_{peak}}{\sqrt{2}}$.
- Ohmic resistors have constant resistance; other devices have resistance that depends on light, temperature, pressure and other quantities.
- Diodes allow significant current to flow in only one direction. Power diodes are used to control current and zener diodes to control voltages.
- Capacitors can produce or manipulate time-varying voltages.



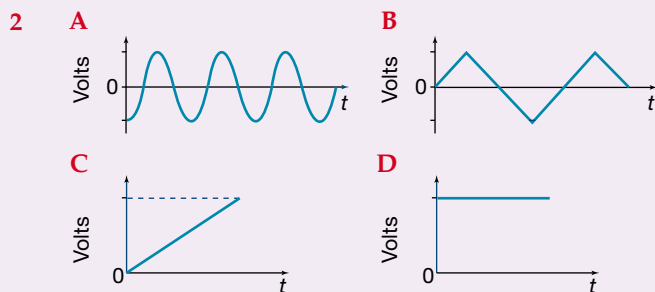
8.1 questions

Principles and practicalities of electronic design

1 The diagram shows a junction in an electric circuit.



- What is the reading on the ammeter?
- What is the direction of the current through the ammeter—towards the junction or away from the junction?



a Which of the graphs A–D best describes the voltage–time relationship for a device that produces:

- i** a steady direct voltage?
- ii** a sinusoidal alternating voltage?
- iii** a non-sinusoidal alternating voltage?
- iv** a steadily increasing direct voltage?

b Which of the graphs would best represent the output of a:

- i** large power station?
- ii** DC generator?
- iii** car battery?

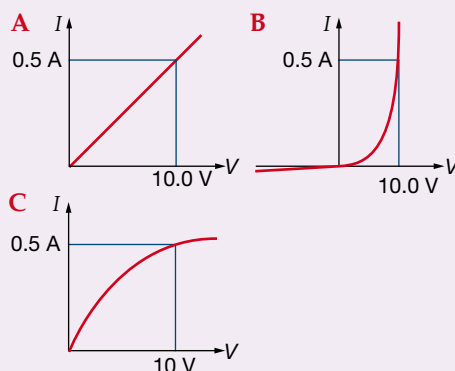
3 The circuit in Figure 8.3 shows a simple transistor amplifier circuit supplied with 6 V. The voltage across the 5 k Ω resistor (i.e. V_{in}) is 2.0 V and the current into the base of the transistor (b) through the 15 k Ω resistor is 100 μ A. The output voltage, V_{out} , is 1.0 V (but no current is flowing in this connection). Use Kirchhoff's laws to answer the following questions.

- a** What is the voltage across the variable resistor, R_T ?
- b** What is the voltage across the load resistor (2.2 k Ω)?
- c** How much current is flowing through the collector terminal (c) of the transistor?
- d** What is the voltage at the base terminal (b) of the transistor?
- e** What is the collector-to-base voltage drop of the transistor?

4 Use the resistor colour code in Figure 8.8 to complete the following table:

	Colour bands in order—first, second, third and tolerance	Value of resistance and tolerance
a	red, black, red and silver	
b	green, orange, yellow and gold	
c		1.5 M Ω + 5%
d		8.2 k Ω + 10%

5 Here are three common graphs of current vs. voltage.



- a** Which graph would represent an ohmic conductor?
- b** Which graph could represent a light globe?
- c** Which graph could represent a power diode?
- d** What is the resistance of each of the conductors when the voltage across it is 10 V?

6 We have studied four types of diodes—power diodes, zener diodes, light-emitting diodes and photodiodes. Which of these would be used:

- a** in a stop light on a car?
- b** in a power supply circuit which converted AC into DC?
- c** to keep the DC voltage produced by such a power supply constant?

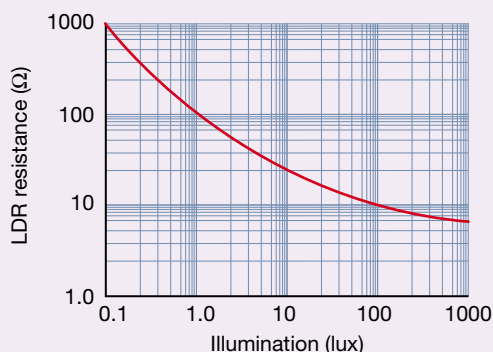
7 The output voltage from a signal generator is displayed on a cathode ray oscilloscope, as shown in the following diagram.



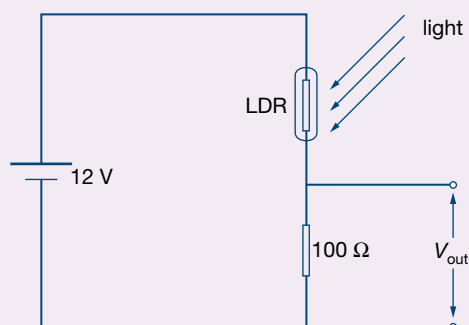
The grid of the CRO screen is divided into squares of 1.0 cm \times 1.0 cm.

- a** The CRO settings are: gain control 5.0 V cm⁻¹, timebase 1.0 ms cm⁻¹.
 - i** What is the value of the peak-to-peak output voltage of the signal generator?
 - ii** What is the RMS output voltage of the signal generator?
 - iii** Determine the frequency of the signal.
- b** The output of the signal generator is adjusted to produce a signal with a peak-to-peak voltage of 20 V and a frequency of 10 kHz. This signal is fed into the CRO and the CRO settings are adjusted to produce an identical trace to that shown above.
 - i** What is the new gain control setting on the CRO?

- ii What is the new time base setting on the CRO?
- c What CRO settings would produce a trace of amplitude 4.0 cm and wavelength 1.0 cm width for a sinusoidal input voltage of 80 V peak-to-peak and frequency 50 Hz?
- 8 A light-dependent resistor (LDR) has the following characteristics. Its resistance depends upon the intensity of the light with which it is illuminated, and decreases as the light intensity increases.



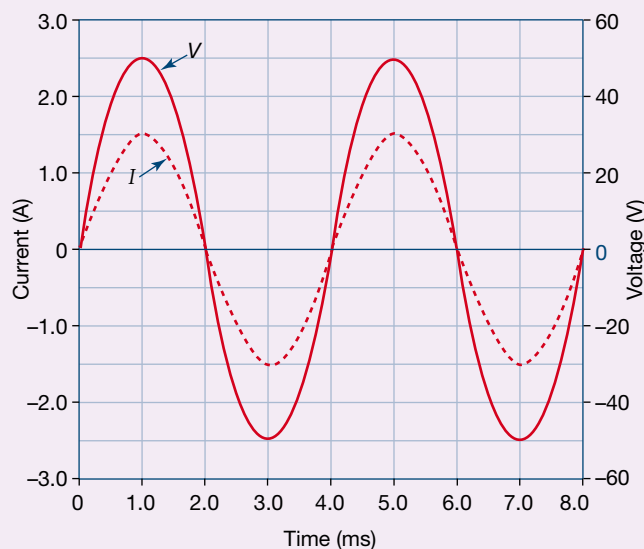
This LDR is connected into the following circuit.



- a What is the resistance of the LDR when the illumination is 100 lux?
- b What is the value of voltage V_{out} when the light intensity is 1.0 lux?
- c What happens to the size of voltage V_{out} as the light intensity increases?
- d A householder wishes to use this circuit to turn on a security light when dusk approaches and the

light intensity drops below a certain value. Output voltage V_{out} is connected to a switch so that when V_{out} rises to 5.0 V the switch operates and turns the security light on. Will the circuit operate the way the householder wishes? If not, what adjustments to the circuit would make it operate correctly?

The following information applies to questions 9 and 10. A student uses a CRO to measure the voltage and current in a loudspeaker that is operating at full power. The resulting voltage–time and current–time graphs are plotted on the same axes, as follows.



- 9 a What is the frequency of the alternating current?
- b What is the peak voltage across the speaker?
- c What is the RMS voltage across the speaker?
- d Calculate the apparent resistance of the speaker. If measured with a multimeter, would this be the value found?
- 10 a What is the peak power the speaker is handling?
- b Calculate the RMS power handled by the speaker.
- c Which value of power do you think best describes the performance of the speaker? Justify your answer.

8.2 Capacitors and time-varying circuits

The electron beam that paints the picture on a TV tube sweeps across the screen 16000 times each second. This is done by a voltage applied to the deflection coils which varies at a frequency of 16000 Hz and is achieved by using a circuit in which a capacitor charges and discharges this many times each second. This is just one example of a time-varying signal. Any electronic device will have many such signals and all are dependent on the properties of capacitors.

To receive the radio signal transmitted from the TV station, a capacitor circuit is again used. In this process, a capacitor and coil circuit *resonates* in time with the desired signal. The ability of a capacitor to generate, or respond to, a time-varying signal is crucial for many functions in electronic circuits. As we shall see, the capacitor also has a crucial role in supplying the steady, non-varying DC voltage needed to power the whole device.

In this section, we look closely at the way in which the time taken to charge and discharge a capacitor can be controlled and used for purposes such as those just mentioned.

A capacitor is a device which stores electric charge. The capacitors used in electronic circuits have two plates, or storage areas, which are placed as close together as possible while still being insulated from each other. When a source of voltage is placed across the plates, opposite charges will flow onto the plates. The closer the plates are, the greater the attraction between these opposite charges. The attraction will result in a larger amount of charge stored on the plates at a given voltage; that is, a greater capacitance.

In fact, the *total* charge on a capacitor is normally zero—there being equal positive and negative charge on the respective plates. When we refer to the charge on a capacitor, we mean the amount of charge that flows onto one plate. (This will be equal to the amount that flowed *off* the other plate, and so the capacitor will be charged to $+Q$ on one plate and $-Q$ on the other.) The amount of charge on the plates depends directly on the voltage across the capacitor, and hence we have the relationship:

$$C = \frac{Q}{V} \text{ or } Q = CV$$

where Q is in coulombs, V in volts and C in farads ($1 \text{ F} = 1 \text{ C/V}$). As 1 C is a huge amount of charge, normal capacitors are rated in smaller units as shown in Table 8.2.

We will concern ourselves with the *rate* at which charge flows onto and off a capacitor. The rate at which a capacitor charges up, or discharges, is the key to many important electronic circuits.

If the leads of an uncharged capacitor are connected to a battery as in Figure 8.17a, electrons flow from the plate at A to the positive terminal of the battery, leaving an overall positive charge due to the loss of electrons (Figure 8.17b). At the same time the negative terminal of the battery provides electrons to the plate at B, giving it a net negative charge. While the capacitor is charging, a current will flow in the circuit. This will continue until the potential difference across the capacitor is equal to that between the terminals of the battery. Once this happens, charge will stop flowing.

When the leads of the charged capacitor are connected, or joined through a circuit, the capacitor will discharge at a rate dependent on the current that will flow in the resistance at the voltage on the capacitor. When both plates of the capacitor reach the same potential, the current ceases. If the capacitor is connected to a different source of voltage, charge will flow on or off until

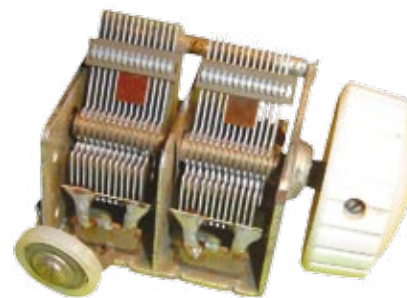


Figure 8.15 The capacitors used to tune older style radios show the basic principle of any capacitor—two sets of plates in close proximity, but insulated from each other. The capacitance of this type is varied by rotating one set of plates into the other.

Physics file

There are two basic categories of capacitors. In one group, the insulator, or dielectric, between the plates is made of materials such as paper, plastic, ceramic, glass and air, or even a vacuum in some cases. These capacitors are often named according to the dielectric material: for example, ceramic capacitors and polystyrene capacitors.

The second category are the electrolytic capacitors, in which a chemical reaction takes place on the surface of the metal plates; for example, the oxidation of aluminium, the product of which becomes the dielectric insulator. Electrolytic capacitors generally store more charge for their size, but are more expensive. They are also polarised; that is, they must be connected in the circuit the right way around or they may be damaged. The symbol for an electrolytic capacitor differs from a non-electrolytic capacitor in order to emphasise its polarity.

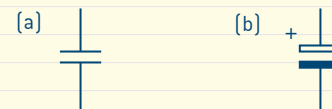


Figure 8.16 (a) A normal capacitor.
(b) An electrolytic capacitor.

Table 8.2 Units used to rate normal capacitors

Unit	Symbol	Value
microfarad	μF	10^{-6} F
nanofarad	nF	10^{-9} F
picofarad	pF	10^{-12} F

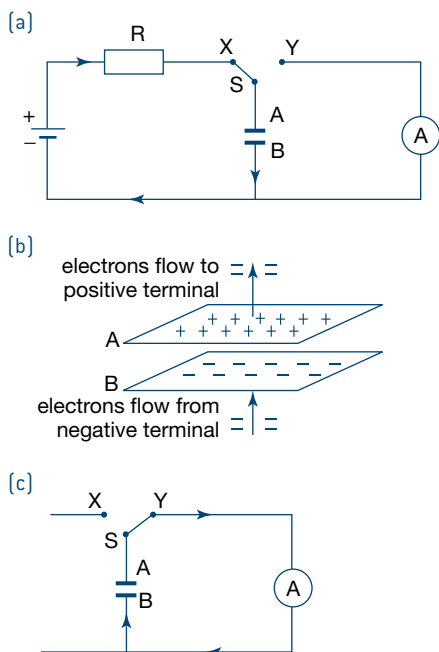


Figure 8.17 A charge–discharge circuit for a capacitor. (a) When the switch is closed to X, charge will flow onto the capacitor. (b) When charging, electrons flow from plate A to the positive terminal of the battery and from the negative terminal to plate B, creating a potential difference between the plates. The battery provides the energy to drive the circuit. (c) When the switch is moved to Y, the battery is disconnected and the capacitor discharges. A current will flow until the plates of the capacitor reach the same potential.

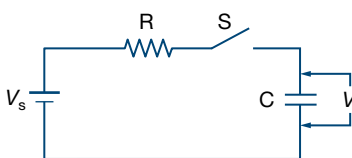


Figure 8.18 When the switch is closed, current will flow at the rate $I = V_s/R$. As the voltage across the capacitor builds up, however, the current will decrease.

the voltage on the capacitor again equals the supplied voltage. A current will exist for a short time while this process occurs.

The time constant, τ

Discharging and charging a capacitor takes time. This is an important consideration in any circuit containing a capacitor. In fact it is often the reason for using the capacitor! For the same supply current, capacitors with a larger capacitance will take longer to charge, since it will take longer to deliver the required amount of charge. The supply current is usually limited by a resistor in series with the capacitor. If this resistance is decreased, the larger current will allow faster charging. The reverse is also true. A large series resistance will reduce the current flow and slow the rate at which the capacitor will charge or discharge.

A capacitor does not charge up at a constant rate. When an uncharged capacitor is initially connected to a source of voltage, current will flow at the rate given by $I = V_s/R$ where V_s is the supply voltage and R the total resistance in the circuit. However, as the charge on the capacitor, and therefore the voltage across it, builds up, the current will decrease as the resultant voltage driving the current through the resistor ($V_s - V_C$) falls. Once the voltage across the capacitor falls to zero, the current will cease. This type of situation, where the rate at which the charge on the capacitor is increasing is dependent on the amount of charge already on the capacitor, leads to an exponential relationship. The closer the voltage across the capacitor gets to the supply voltage, the smaller the current becomes. The mathematics of this is described in the Physics file on page 295, but all we really need to know is that the voltage builds up with time in the way shown in the graph in Figure 8.19 and that after a time given by $\tau = RC$, the capacitor is about 63% charged. This time, τ (the Greek letter tau), is referred to as the *time constant* for the circuit. It is simply the product of the total resistance in the charging circuit and the value of the capacitance.



The time it takes for a capacitor to become charged depends on its capacitance, C (F), and the resistance, R (Ω), in the rest of the circuit. The **TIME CONSTANT**, $\tau = RC$, represents the time (s) to reach 63% of the full charge.

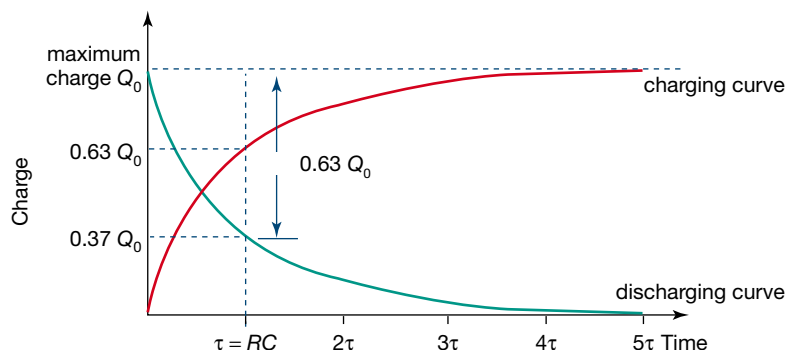


Figure 8.19 A capacitor will charge or discharge in an exponential fashion. A capacitor will charge or discharge to 63% of its maximum charge after time $\tau = RC$. After a time equal to 5τ , a capacitor is taken to be fully charged or discharged.



PRACTICAL ACTIVITY 28

Determining the time constant in an R–C circuit

Mathematically speaking, we could say that a capacitor takes forever to become fully charged (as the graph asymptotes to the maximum charge line), but in practice, electronic engineers regard five time constants (5τ) as being enough to 'fully charge' the capacitor. (You might like to show that at this time it will be over 99% charged.) Because the mathematics of discharging the capacitor is basically the same, the time constant is the same whether charging or discharging. That is, it will take one time constant to discharge the capacitor to 37% of its initial voltage.

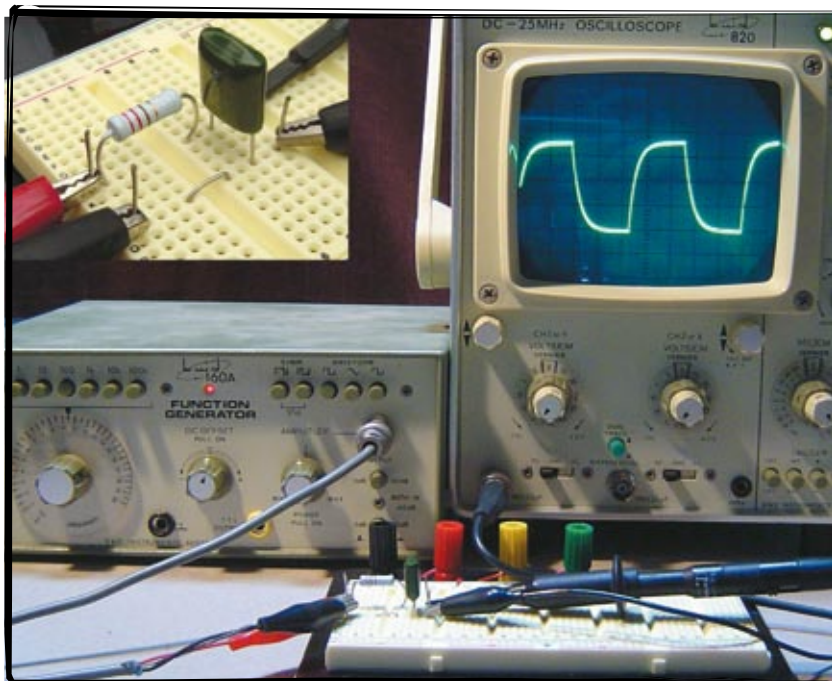
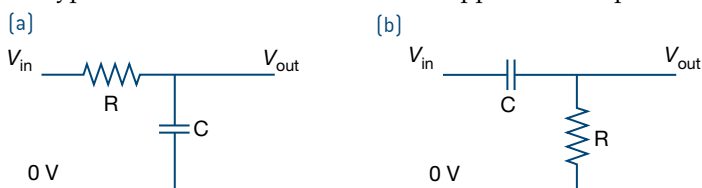


Figure 8.20 The charge and discharge curves for a capacitor can be displayed on a CRO or computer screen. This enables a measurement of the time constant and hence a determination of the capacitance, provided the resistance in the circuit is known. The insert shows the circuit detail.

R-C circuits

The effect that capacitors have in electronic circuits can sometimes be rather confusing. However, most uses of capacitors in practical situations can be characterised by one of the two simple circuits illustrated in Figure 8.21.

In Figure 8.21a, the capacitor is used to smooth out fluctuations in the input voltage (V_{in}). In this configuration, current through the resistor will charge the capacitor when V_{in} is higher than V_{out} , but if V_{in} is lower than V_{out} the capacitor tends to maintain V_{out} at the higher voltage. If V_{in} happens to be an oscillating voltage, such as an AC voltage, or the signal voltage from a sound source in an amplifier, this circuit will tend to smooth it out; that is, it will reduce the amplitude of the signal. The larger the capacitor and resistor, the longer the time constant and the more effective will be the smoothing. As we shall see, this type of circuit is used to smooth out ripples in a DC power supply.



Physics file

The mathematics of the charging of a capacitor is similar to that of radioactive decay or of population explosions in biology, where the rate of change of a quantity depends on the instantaneous value of that quantity. (In the first two, the rate decreases, whereas in a population explosion the rate increases.) The rate at which the capacitor charges is equal to the current flowing in the circuit. This is given by:

$$I = \frac{V_s - V_c}{R}$$

But as $V_c = \frac{Q}{C}$ and $I = \frac{dQ}{dt}$ this equation can be written as:

$$\frac{dQ}{dt} + \frac{Q}{RC} = \frac{V_s}{R}$$

Mathematicians will recognise that the solution of this differential equation produces an exponential relationship of the form $Q = CV_s(1 - e^{-t/RC})$. As t increases from zero to very large, the $e^{-t/RC}$ term decreases from 1 to zero, leading to the graph of charge shown in Figure 8.19. When $t = RC$ the charge has reached $(1 - e^{-1})$ or 0.63 of its final value. This value is said to be the *time constant* or τ for the circuit.

Physics file

The time taken to charge or discharge a capacitor becomes a particular problem in computer and digital information circuits. Any circuit wiring will have some unavoidable capacitance. If digital information is being processed at a frequency of 1000 MHz (1 GHz) the time for one cycle is one-billionth of a second, and this *circuit capacitance* needs to charge or discharge during this incredibly short time. Integrated circuits (ICs) attempt to overcome this problem by having tiny connections so that they have a very low capacitance.

Figure 8.21 Most uses of a capacitor in a circuit can be characterised by one of these arrangements. (a) A low-pass filter and (b) a high-pass filter (discussed in the Physics in action on page 296).

Low-pass and high-pass filters

These two circuits are the basis for important components of many electronic devices that allow or block signals of certain frequencies. They are known as low-pass or high-pass filters. If the time for the variations of V_{in} is long compared with the time constant of the circuit, the capacitor will be able to charge or discharge and so V_{out} will 'follow' V_{in} . For this reason, the circuit configuration in Figure 8.21a is often called a *low-pass filter*. A low-frequency oscillating voltage at V_{in} will pass through the circuit, but a higher frequency voltage will be blocked by the circuit; that is, the signal will be smoothed out.

On the other hand, Figure 8.21b has the opposite effect and so is often called a *high-pass filter*. Let us see why. It is important to remember that it takes time to change the potential *difference* across the terminals of a capacitor. Let us imagine that V_{in} is at a steady 5 V and V_{out} is at 0 V (as it is connected to the 0 V line via the resistor). There is, therefore, a difference of 5 V across the capacitor. Now a sharp rise of 2 V in V_{in} occurs, bringing V_{in} to 7 V. At this point the voltage difference across the capacitor is still 5 V and so V_{out} immediately rises to 2 V. This will result in a current through the resistor which will gradually increase the charge on the capacitor until there is a 7 V difference across it. This lowers V_{out} which eventually drops back to zero. In other words, rapid changes in voltages will be passed through the circuit, but not slow ones; hence the term *high (frequency) pass filter*.

These types of circuits can be used to filter the sound signals sent to the speakers in a hi-fi sound system. The tweeter, or high-frequency speaker, requires only the high-frequency signals and so a high-pass filter (Figure 8.21b) is used. The woofer, or bass speaker, only requires the low frequencies and so the signal to it is passed through a low-pass filter (Figure 8.21a). In practice, the filters used in good hi-fi systems can also contain inductors (coils which have electromagnetic properties) in order to achieve sharper cut-off frequencies. Figure 8.22 shows the effects of the two circuits on oscillating signals of various frequencies.

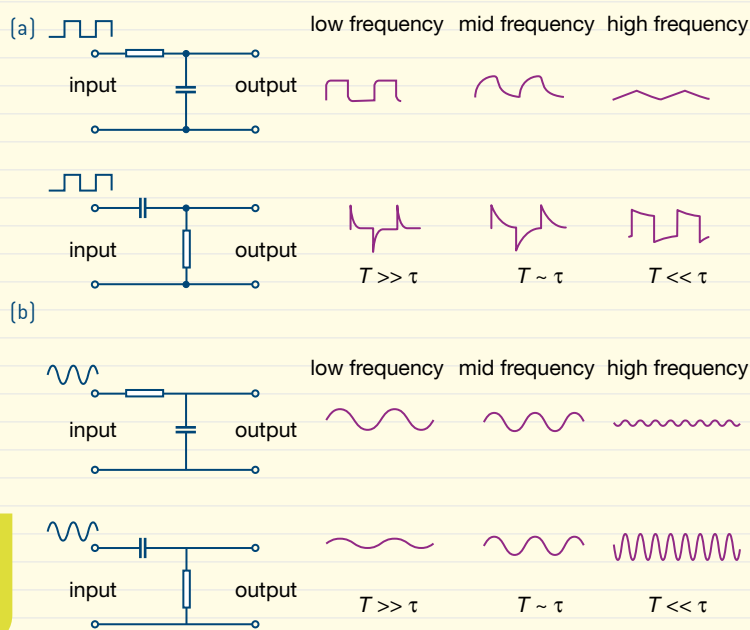


Figure 8.22 (a) The effect of the two types of R-C circuits on a sharply changing voltage (usually called a square wave) with a low, mid and high frequency relative to the time constant of the circuit. (b) The effect on sine wave signals of similar frequencies.



8.2 summary

Capacitors and time-varying circuits

- Capacitors store an amount of charge which is proportional to the voltage and capacitance:

$$C = \frac{Q}{V}$$

- The time taken to charge a capacitor depends on the capacitance and the resistance in the circuit. This is known as the time constant, τ :

$$\tau = RC$$

- After one time constant, τ , the capacitor will be about 63% charged (or discharged). After 5τ it is regarded as fully charged (or discharged).
- A sudden change in potential of one terminal of a capacitor will result in the same sudden change on the other terminal until the capacitor has time to adjust to the change.
- A simple capacitance-resistance circuit can be used to modify an AC signal.

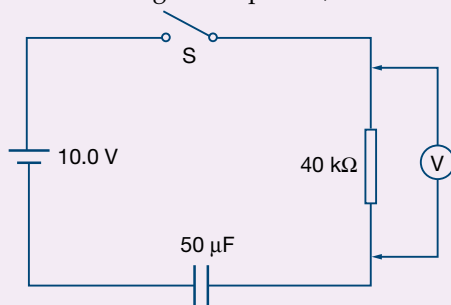


8.2 questions

Capacitors and time-varying circuits

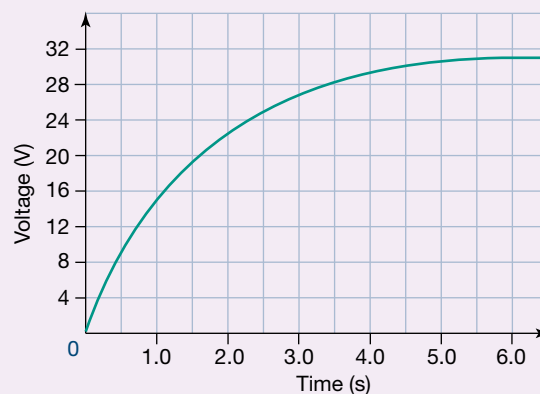
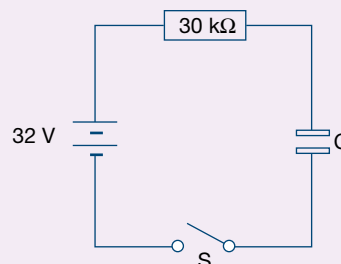
- Which of the following is equivalent to one farad?
A one coulomb.volt
B one volt per coulomb
C one coulomb per volt
D one coulomb per joule
E one joule per coulomb
- A potential difference V is placed across a capacitance C resulting in a charge Q on the plates of the capacitor. Complete the following table by calculating the missing quantity in each case. (See Table 8.2 for the values of the prefixes used.)

	C	V	Q
a	0.5 F	100 V	
b	10 μF	400 V	
c		250 V	20 nC
d		90 V	9.0 pC
e	0.5 μF		0.01 C
- Define the time constant of a capacitor–resistor circuit and describe its meaning.
 - Calculate the time constant for each of the following combinations of resistance R and capacitance C :
 - $R = 1.0 \text{ k}\Omega$, $C = 10 \text{ pF}$
 - $R = 80 \Omega$, $C = 5.0 \text{ mF}$
 - $R = 2.0 \text{ M}\Omega$, $C = 0.10 \text{ pF}$
- A capacitor has $\tau = 10 \text{ ms}$. Which option gives the best estimate of the time it would take the capacitor to effectively fully charge or discharge?
A 6.3 ms **B** 20 ms **C** 50 ms **D** 63 ms
- A 50 μF capacitor is charged up using the circuit shown here. To charge the capacitor, switch S is closed.



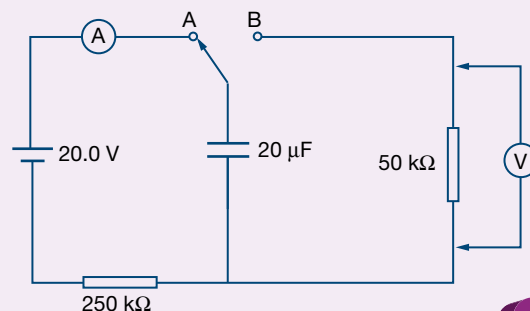
- What is the time constant for this circuit?
- How long will it take for the capacitor to reach its full charge?
- After the capacitor is fully charged, what is the reading on the voltmeter?
- What would the reading on the voltmeter be 2.0 s after switch S is closed and the charging process commences?

The following information applies to questions 6 and 7. This circuit is used to produce the charge curve shown.



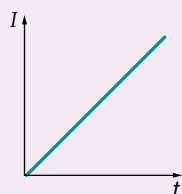
- What is the time constant for the circuit?
 - What is the value of the capacitor?
 - How much charge is stored in the capacitor when it is fully charged?
- 7 While the capacitor is charging, which of the following is true?
- The resistor is converting energy and the capacitor is storing energy.
 - The resistor is storing energy and the capacitor is converting energy.
 - Both the resistor and the capacitor are converting energy.

The following information applies to questions 8 and 9. A capacitor is charged and then discharged using the following circuit. Initially the switch is in position A so that the capacitor is charged.

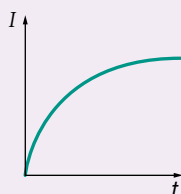


- 8 a How long does it take for the capacitor to reach a full charge?
 b After one time constant what is the potential difference across the capacitor?
 c At this time (i.e. after one time constant), what is the reading on the ammeter?
 d Which of the following graphs would best represent the current through the $250\text{ k}\Omega$ resistor versus time during the charging process?

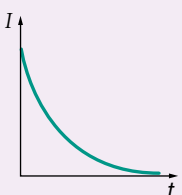
A



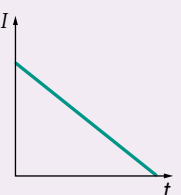
B



C

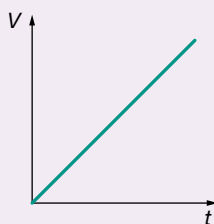


D

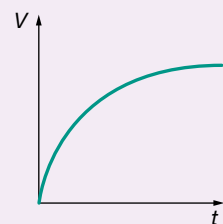


- 9 The switch is now moved to position B so that the capacitor is discharged.
 a How long does it take for the capacitor to fully discharge?
 b After one time constant what is the reading on the voltmeter?
 c Which of the following graphs shows the potential difference across the resistor versus time as the capacitor discharges?

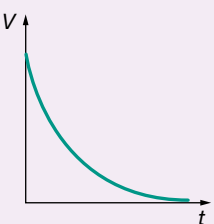
A



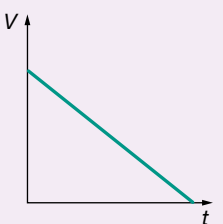
B



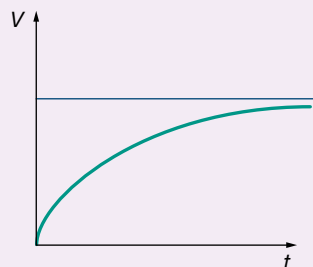
C



D



- 10 a A capacitor is charged up in a circuit consisting of a 12 V battery, the capacitor and a $10\text{ k}\Omega$ resistor. The following graph shows the voltage changes during the charging process.



How would the graph change in shape if the resistor was changed to one of $5\text{ k}\Omega$? (Show your answer by copying this graph and then drawing the new graph in a different colour. Explain why you made the changes you made.)

- b The potential difference or voltage across a capacitor which is discharging follows a mathematical rule of the form:

$$V = V_0 e^{-t/\tau}$$

where V is the voltage at time t , V_0 is the voltage at the start, and τ is the time constant. If $V_0 = 10\text{ V}$ and $\tau = 2\text{ s}$, then the equation would become $V = 10e^{-t/2}$.

Use your calculator to complete the following table, and then plot a graph of V versus t to investigate its shape.

t	$t/2$	$e^{-t/2}$	$V = 10e^{-t/2}$
0.0			
1.0			
2.0			
3.0			
4.0			
6.0			
8.0			
10.0			

8.3 Rectification and power supplies

The electricity supplied to our homes is AC with an RMS voltage of 240 V, but almost all of the electronic devices in our homes require low-voltage DC power. For example, the control panel on a microwave oven or heating system would typically operate at between 5 and 12 V DC. All radio and television sets as well as all computers and their peripherals require low-voltage DC. It is not practical or economical to run these circuits from batteries, so power supplies are built into these appliances to convert the relatively high AC mains voltage into low-voltage DC.

Transformers are used to step down (decrease) or step up (increase) a supplied voltage to the voltage required for a particular application. They consist of two coils arranged so that the alternating current in the primary coil induces an alternating current in the secondary coil. Electromagnetic induction theory (see Chapter 10) tells us that, in an efficient transformer, the ratio of the voltage across the primary coil to the voltage across the secondary coil equals the ratio of the number of turns in each coil, and so:

$$\frac{V_p}{V_s} = \frac{N_p}{N_s}$$

Provided the transformer itself consumes little of the power that passes through it (and this is usually the case) we know that the power into the primary coil is equal to the power provided by the secondary coil, and hence $V_p I_p = V_s I_s$. Putting these equations together, we can see that, for an ideal transformer:

$$\frac{V_p}{V_s} = \frac{N_p}{N_s} = \frac{I_p}{I_s}$$

In our practical work in this study, we will assume that you have access to a low-voltage AC source from a commercial transformer or laboratory power supply. It is dangerous, and illegal, to tamper with mains voltage fittings. The rest of this section deals with ways to produce a steady, low-voltage DC power supply from a source of low-voltage AC.

Rectifiers

Once a high-voltage supply has been reduced to a suitably low voltage by a transformer, the negative section of the AC cycle needs to be blocked or altered so that there is a flow of current in one direction only, as is the case with direct current. This process is called *rectification*. (*Rectify* means to correct, and in this context AC is being 'corrected' to DC.) Rectification can be half-wave or full-wave. Half-wave rectification simply blocks the negative part of the AC cycle; full-wave rectification converts the negative cycle into a positive one.

Half-wave rectification

The simplest rectifying circuit is a half-wave rectifier which uses a single diode to ensure that current flows in only one direction. An arrangement like the one shown in Figure 8.24 allows only half of each cycle from the AC supply to flow through the circuit. (Can you see why? The diode only conducts when it is forward biased. On the negative half of the cycle it is reversed biased.)

If the voltage across the resistor in the circuit is monitored by a CRO, as no current flows in the negative half of the AC cycle, the output voltage is

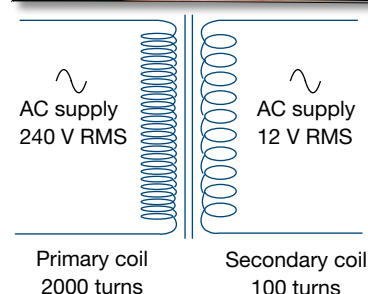
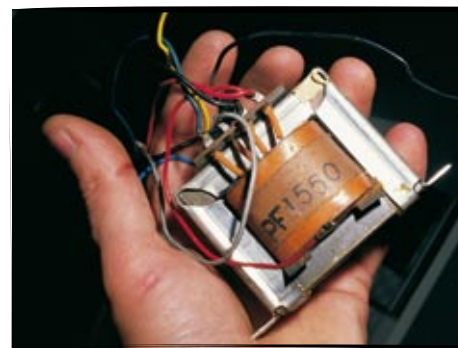


Figure 8.23 In a transformer, the ratio of the supply voltage [across the primary coil] to the induced voltage [across the secondary coil] equals the ratio of the number of turns in the respective coils. In this example, which is typical of a transformer in a household appliance, $240 \text{ V}/12 \text{ V} = 2000 \text{ turns}/100 \text{ turns} = 20$.

zero during that time. This means that the voltage across the load resistance will vary in value between zero and the peak voltage value of the original AC supply, but for half the cycle the output will be zero. Although the potential difference varies, it is still considered to be direct because it acts only in one direction. But this circuit is actually wasting half the available power since it is effectively turned off during half of every cycle. So although it is easy to construct, this circuit is of limited use.

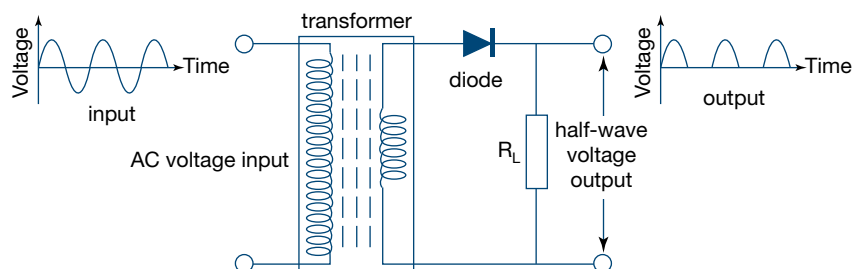


Figure 8.24 A half-wave rectifier is a simple circuit using a single diode. It blocks the negative half of each AC cycle.

Full-wave rectification

An ideal rectification circuit will ensure that all the available energy from an AC supply is converted into DC. Rather than simply blocking the negative part of the cycle, it is better to reverse it so that it is in the same direction as the positive part of the cycle. This is called *full-wave rectification*.

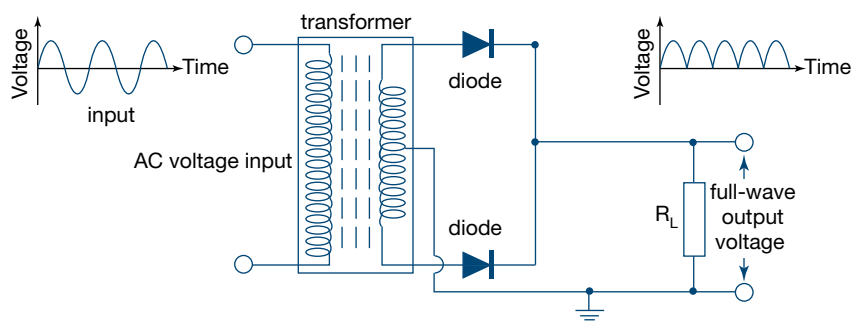


Figure 8.25 Using two diodes connected to a centre-tapped transformer voltage, both the positive and negative cycles of an AC voltage can be used, resulting in full-wave rectification.

One way to achieve full-wave rectification is to use a secondary coil in the transformer which has a centre-tap; that is, a connection to the centre of the coil (Figure 8.25). Two diodes are connected at opposite ends of the secondary coil but are both connected to one end of the load, as shown. The other end of the load is connected to the centre-tap. This means that for each half of the AC cycle, current flows through either one diode or the other, depending on which end of the secondary coil is positive. This results in a current through the load which is always in the same direction (down in this diagram). While this is a DC current, it is not a steady DC current, as can be seen from the graph in Figure 8.25. However, current now flows in both halves of the AC cycle, varying between zero and the original AC peak current value.

A centre-tapped transformer is more expensive than a simple complete coil transformer to construct. A cheaper method for achieving full-wave

rectification makes use of four diodes, creating what is called a *diode bridge* (Figure 8.26). The current paths are a little more difficult to follow than those in a two-diode system, but the overall effect of a bridge rectifier is the same. The AC current is fully rectified, using each part of the cycle. With a bridge rectifier there is no need for the transformer to have a central wire, or tap, to complete the circuit to the load, so a diode bridge enables the use of cheaper transformers. The four-diode bridge, while seemingly complicated, is actually quite a simple arrangement and can be constructed from four diodes or purchased as a ready-made mass-produced single component.

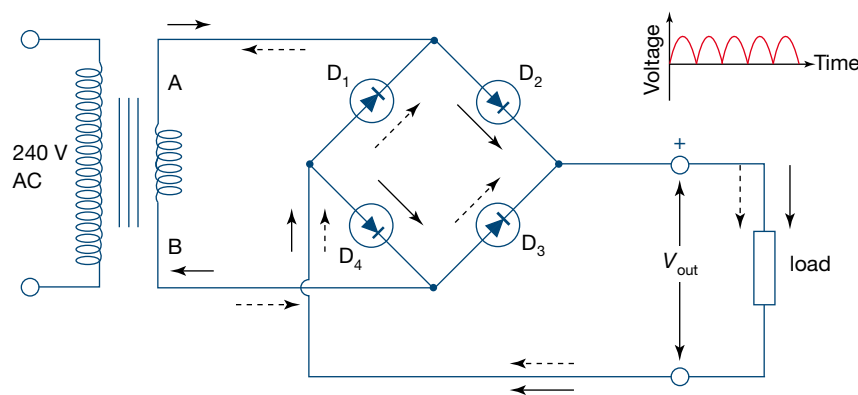


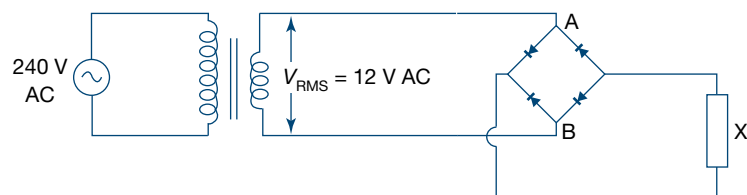
Figure 8.26 The economical bridge rectifier circuit gives full-wave rectification without the need for an expensive centre-tapped transformer. The four-diode bridge is cheap and easy to mass-produce. The solid and broken arrows show the current path in the two parts of the AC cycle.

To see how the diode bridge rectifies the AC, first imagine the top of the transformer (A) is positive with respect to the other end (B). In this situation, shown by the solid arrows, diode D_2 will conduct and allow a positive current through the load. This current returns to the transformer via D_4 . The other two diodes are reversed biased in this part of the cycle and so do not conduct. In the other part of the cycle when the lower end of the transformer is positive, diodes D_1 and D_3 conduct (as shown by the broken arrows), but the current again flows in the same direction through the load.

Full-wave rectification is closer to a smooth DC current than half-wave rectification, but for many real applications a constant voltage output like that produced by a battery is needed. For a rectified supply to be useful, the variations in current and voltage need to be smoothed out to give a constant value. We will look at how this can be achieved shortly.

Worked example 8.3A

In an electronic clock, the 240 V AC mains supply is converted to 12 V AC using a step-down transformer. The current then passes through a rectifier, illustrated below. The remainder of the clock's circuitry can be regarded as a single-load resistance, shown as X in the diagram.

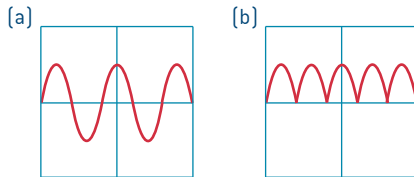


- a** What will be the ratio of the number of turns in the primary coil to the number of turns in the secondary coil in the transformer?

- b** If a CRO is connected across points A and B, what would the trace look like? Illustrate your answer.
- c** If the CRO is connected across load X, what would the new trace look like?

Solution

- a** The ratio of turns is the same as the ratio of the voltages:
i.e. $\frac{V_p}{V_s} = \frac{N_p}{N_s} = \frac{240}{12} = 20$
- b** Points A and B receive current directly from the transformer, without it passing through any of the diodes. The current will therefore still be AC. The trace would be as in diagram [a].
- c** The rectifier is a bridge circuit which would fully rectify the current. The voltage in the load would now be fully rectified and the trace would be as in diagram [b].



Smoothing out the power

As was suggested in section 8.2, a capacitor is needed to smooth out the oscillating voltage from our full-wave rectifier. If a capacitor is connected across the output of a rectifier circuit, it will charge up whenever the output voltage is greater than the potential difference across the capacitor. So in each cycle, the capacitor is charged to the peak output voltage. As the output voltage drops, the capacitor will start to discharge through the load. Given large enough capacitor and resistance values (i.e. a relatively large time constant), the capacitor will maintain a voltage fairly close to the peak output value, thus 'filling in the gap', or *smoothing* the output cycles from the diode bridge.

The degree of smoothing will depend on the size of the capacitor and the current drawn by the load. For a circuit of given resistance, a larger capacitor with a longer discharge time will hold the voltage closer to its peak value for longer and so create a steadier output. After smoothing, there will still be some small variation, called a ripple voltage. The size of this *ripple voltage* is a measure of how closely the output resembles the steady voltage from a battery: the smaller the ripple the better. The equivalent DC voltage will lie somewhere between the minimum and maximum values of this smoothed, or filtered, output.

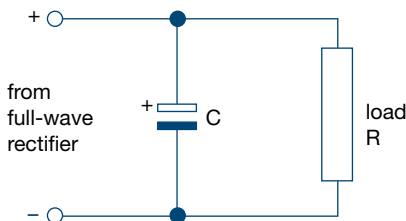


Figure 8.27 The addition of a capacitor to the output of a full-wave rectifier smooths out the voltage pulses.

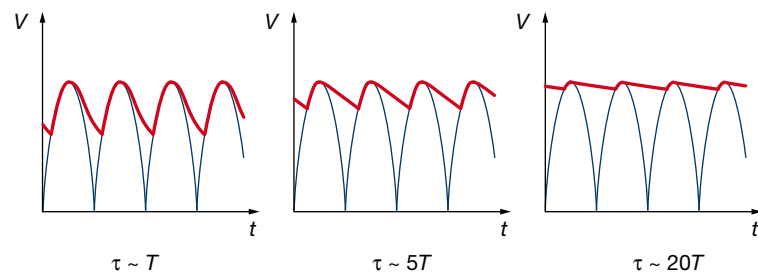


Figure 8.28 The addition of a capacitor to the output of a full-wave rectifier smooths out the peaks. The greater the value of the time constant ($\tau = RC$), the better the amount of smoothing. There will, however, always be a certain amount of ripple remaining in the voltage to the load.

In order to ensure a very low ripple voltage, it is necessary to use a large value capacitor. However, capacitors are relatively expensive and so it is not good design to use one that is unnecessarily large. How then do we determine the optimum value for the capacitor? The frequency of the cycles from the rectifier is 100 Hz, as there are two cycles in every 50 Hz AC cycle.

Thus the period is 0.01 s or 10 ms. Clearly, in order to provide a reasonably smooth output, the time constant for the circuit needs to be much longer than this. Just how much longer depends on the application.

The time constant depends on the values of the capacitor and resistor in the circuit. While the capacitance is fixed, the load resistance will not normally be constant. Consider a radio, or other sound amplifier circuit, for example. If the volume of the radio is turned down, there will be a relatively low current drawn from the power supply, but when turned up with loud music playing, the current may increase to many times the minimum current. So in designing a power supply circuit, it is important to know not just the average current drawn, but the peak value of the current to be supplied. As far as the power supply is concerned, this fluctuating current load looks like a varying resistance: a high current corresponding to a low resistance. The effective resistance can be calculated simply from the voltage supplied and the current drawn. For example, a load requiring 12 V and a peak current of 100 mA would appear to be a load resistance of:

$$R = \frac{V}{I} = \frac{12}{0.1} = 120 \, \Omega$$

If we assume that a circuit with a time constant of, say, five times the period of the power supply ripple will be satisfactory, we can determine the value of capacitance required. As the period of the rectifier output is 10 ms, the time constant required is $\tau = 5 \times 10 = 50$ ms, or 0.05 s, and the capacitance needs to be:

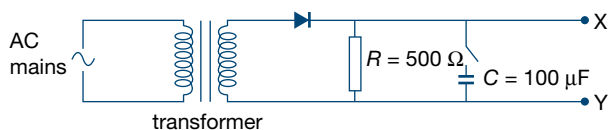
$$C = \frac{\tau}{R} = \frac{0.05}{120} = 0.00042 \, \text{F} = 420 \, \mu\text{F}$$

(Be careful with the unit multipliers in these types of calculations!) Actually, as is shown in the adjacent Physics file, this would still result in a ripple voltage of nearly 20%. For some applications, a significant ripple voltage may not matter. If, for example, the purpose of the circuit is simply to light an LED or run a motor, small fluctuations in the supply voltage will not be noticeable. On the other hand, if the power supply is to run an audio amplifier, a ripple voltage at 100 Hz may well be heard as a very annoying, low-frequency hum.

To reduce the ripple to less than 1%, it is necessary to choose a time constant about 100 times the period of the power supply (i.e. τ would be 1 s). In our 12 V, 100 mA circuit, that would require an 8400 μF capacitor. You can see that for the power supplies required in circuits that use significant current (and our 100 mA is only a very modest amount), very large capacitors are needed. If you have noticed that your radio does not come on immediately when you press the switch, the delay may well be due to the time taken for the large power supply capacitors to charge up.

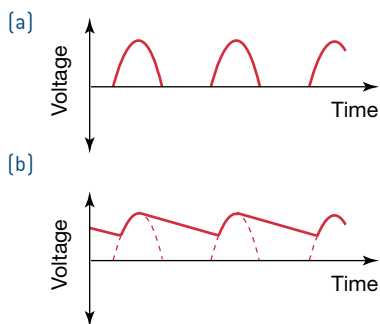
Worked example 8.3B

The circuit shown in the following diagram is used to demonstrate half-wave rectification. The capacitor can be connected by closing the switch to demonstrate smoothing of the rectified voltage.



Physics file

Earlier we showed that as a capacitor *charges* from a power supply of V_S volts the charge on it is given by the expression $Q = CV_S(1 - e^{-t/RC})$. The expression for the charge remaining as the capacitor is *discharged* through a resistor is found, in a similar way, to be $Q = CV_S e^{-t/RC}$. When t is zero, the $e^{-t/RC}$ term is 1, and as t increases this expression decreases to zero. A little work with a calculator will show that after one-fifth of a time constant, (or $t/RC = 0.2$) the charge (and hence voltage) is already down to 82%, so that if a time constant five times the cycle period was chosen the ripple voltage would be nearly 20%. You can show that in order to reduce the ripple to 5%, the time constant needs to be about 20 times the cycle period. To reduce it to 1%, the time constant required is 100 times the cycle period.



- a Sketch the rectified voltage across the output X–Y with the switch open.
- b Describe the change in the output when the capacitor is connected to the circuit by closing the switch.
- c Find the time constant for this combination of resistor and capacitor and hence sketch the voltage at X–Y with the capacitor connected in the circuit.
- d How could the circuit be changed in order to achieve a smoother output voltage?

Solution

- a With only one diode, the circuit provides half-wave rectification. The output from the rectifier will resemble graph (a).
- b The capacitor will smooth the output from the rectifier, charging on the 'up' side of the output pulses, and then discharging and thus maintaining a current flow as the output potential drops again. The amount of smoothing depends on the values of the capacitor and resistor.
- c From the diagram of the circuit, $R = 500\ \Omega$, $C = 100\ \mu\text{F}$
 The time constant $\tau = RC = 500(100 \times 10^{-6}) = 5.0 \times 10^{-2}\ \text{s}$ or 50 ms.
 This is $2\frac{1}{2}$ times the period of the 50 Hz mains pulses and so the voltage will fall significantly between each pulse. Graph (b) shows the possible resulting output.
- d Smoothing of the rectified output is improved by decreasing the rate of discharge, i.e. by increasing the time constant τ . As the equation suggests, the discharge time, and hence smoothing, is influenced by both resistance and capacitance. Better results can be achieved by increasing the resistance and/or capacitance.

You may have noticed in Worked example 8.3B that it was assumed that the capacitor charged up rapidly to the peak voltage with the 'up' side of the pulse from the rectifier. As with the discharge cycle, the time constant is determined by the values of the capacitance and resistance involved. The resistance in this instance, however, is that of the transformer coil and the diode. Both of these will normally be very low compared with the load resistance and so the time constant for the 'charge' part of the cycle will be very much shorter than that of the 'discharge' cycle. If, in fact, the rectifier circuit contained significant resistance, the voltage across the capacitor might not reach the peak voltage. (Notice that in the discharge cycle, the diode is not conducting and therefore effectively has an infinite resistance.)

Regulating the voltage

As we have just seen, very large capacitors are needed to achieve a really smooth DC output from a rectifier circuit. Large capacitors are both physically large and relatively expensive, and neither of these attributes is desirable in today's world of miniaturised electronics. Not surprisingly then, electronic engineers have found other ways of smoothing the power supply voltage. *Voltage regulators* have the effect of cutting the ripple from the output of a rectifier circuit.

Before looking at regulators themselves, we will consider a simpler device which can be used for the same purpose, and which is indeed an essential component of a regulator: the zener diode. This type of diode has a relatively low, but predictable, reverse-breakdown voltage. It is used 'backwards' in a circuit; that is, in such a way that it is reverse biased. If the supply voltage exceeds the breakdown voltage, it will begin to conduct. Remember that zener diodes are designed to be used this way and are not damaged by the

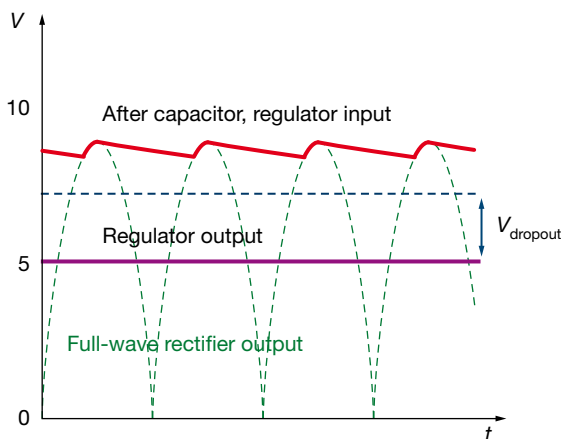


Figure 8.29 A voltage regulator has the effect of removing the ripple from the output of a rectifier circuit. V_{dropout} is the minimum voltage drop required between input and output of the regulator and is usually about 2 V.

reverse current provided they are used within their specified ratings. Let us look at how they can be used to regulate the voltage supplied to a load.

Figure 8.30 shows one way to reduce the supply voltage to the appropriate level for a particular device or application. The dropping resistor is set so that there is an appropriate voltage drop across it, leaving the correct voltage drop across, in this case, the light bulb. The problem with this circuit is that the resistor has to be adjusted for the amount of current that is drawn by the light bulb. If a different 6 V light bulb was placed in the circuit, it might either burn out or operate at lower brightness, depending on the current it required. The zener diode can be used to overcome this problem.

By placing an appropriate zener diode in the circuit, as shown in Figure 8.31, the voltage across the load (for example, the 6 V light bulb) is stabilised. If the load draws too little current, the voltage, V_L will rise and the zener will start to conduct. This will result in a higher current through R_d which in turn will lower V_L . An equilibrium will be set up so that the zener conducts just sufficiently to maintain a constant, *regulated*, voltage across the load. Clearly there are limitations that will restrict the use of this circuit. For one thing, the current flowing through the dropping resistor, as well as through the zener diode, will result in power lost as wasted heat. If too much power is lost this way either the dropping resistor or the zener diode may burn out. As well, a compromise must be reached between the power wasted and the ability of the circuit to cope with higher currents, because if resistance is too high, the voltage will fall too far when the current increases. Worked example 8.3C illustrates this point.

Worked example 8.3C

A 1N4737 zener diode has a 7.5 V zener voltage and is rated at 1 W power dissipation. It is to be used in the circuit shown at right to regulate the voltage to a small radio circuit which requires a constant voltage of 7.5 V. The supply voltage can vary between 9 V and 12 V.

- What is the maximum current that can pass through the diode without exceeding the power rating?
- What is the maximum current that can be drawn by the radio in this circuit?
- What is the minimum value of the dropping resistor, R_d , that can be used in this circuit?
- What could happen if a resistor of lower value was used?
- Why not use a higher value resistor?

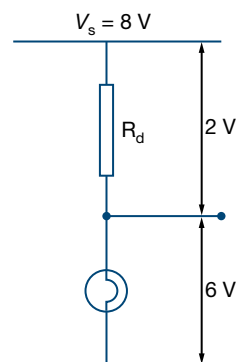


Figure 8.30 This circuit shows one way to supply the correct voltage to a light bulb from a higher voltage supply (8 V). The 'dropping' resistance R_d is adjusted so that there is a 2 V drop across it, leaving the light bulb with the correct voltage.

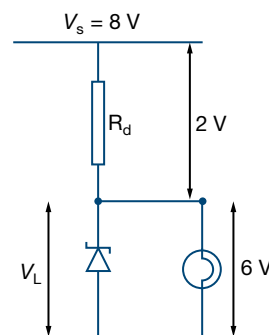
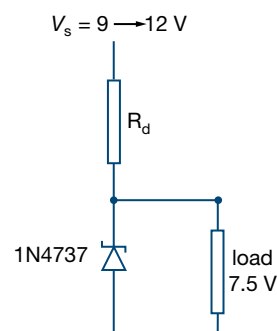


Figure 8.31 A 6 V zener diode is used to regulate the voltage supplied to the light bulb. If the voltage across the bulb rises, due to an increase in V_s or the use of a different 6 V light bulb, the zener will conduct, causing a larger potential drop across R_d .



- f** If the power supply falls to 9 V, what is the maximum current available to the radio?
- g** If the power supply is at 10 V and the radio is drawing 40 mA, how much power is being wasted by the circuit?

Solution

- a** The voltage across the diode will be 7.5 V and therefore the maximum current at 1 W will be given by:

$$I = \frac{P}{V} = \frac{1}{7.5} = 0.13 \text{ A or } 130 \text{ mA}$$

- b** The maximum current is limited to the maximum current that can pass through the diode because if the radio is drawing no current, all of the current passing through the dropping resistor must pass through the diode instead. So the radio must draw less than 130 mA.

- c** The minimum safe value of R_d is determined by the maximum voltage drop required across it ($12 - 7.5 = 4.5 \text{ V}$) and the maximum allowable current through it (130 mA). Thus:

$$R_d = \frac{V}{I} = \frac{4.5}{0.13} = 35 \Omega$$

- d** If, say, a 30Ω resistor was used and the radio was switched off, the possible current through R_d and the diode would be given by $I = \frac{4.5}{30} = 150 \text{ mA}$, which would exceed the 1 W power rating of the diode.

- e** A higher value resistor will result in less than 130 mA being available to the radio, as the drop across the resistor would be greater than the 4.5 V required.

- f** Given the 35Ω dropping resistor, and the voltage drop across it of 1.5 V ($9 - 7.5$), the maximum current available will be:

$$I = \frac{V}{R} = \frac{1.5}{35} = 0.043 \text{ A or } 43 \text{ mA}$$

- g** In this situation the current through R_d will be:

$$I = \frac{V}{R} = \frac{2.5}{35} = 71 \text{ mA}$$

Of this 71 mA, 40 mA will be used by the radio and the other 31 mA will pass through the diode. The power lost in the dropping resistor will be $2.5 \times 0.071 = 0.18 \text{ W}$, and that in the diode will be $7.5 \times 0.031 = 0.23 \text{ W}$, a total of 410 mW. This can be compared with the actual power used by the radio: $7.5 \times 0.04 = 300 \text{ mW}$. Clearly there is considerable power wasted in this arrangement!

Voltage regulators

Fortunately, electronic engineers have come up with a better solution to the problems of ripple voltage and fluctuating power supplies! It is called the *voltage regulator*. It is a small three-terminal integrated circuit which takes the varying input from a rectified power supply and produces a steady DC preset voltage at the output. Needless to say, voltage regulators were a great boon to electronic engineers and hobbyists! The input voltage normally needs to be at least 2–3 V higher than the desired output in order to obtain the specified voltage. Figure 8.32 shows the use of a 7805 voltage regulator. The three terminals are:

- 1 the input from the power supply
- 2 the common or ground (0 V)
- 3 the output, the smoothed and regulated voltage ready for use.

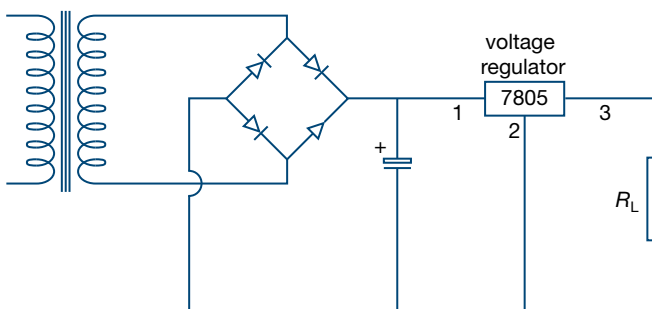


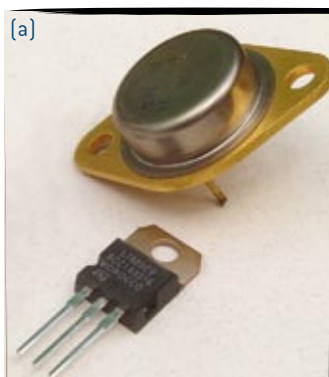
Figure 8.32 A typical power supply circuit utilising a 7805 voltage regulator.

The specifications for a typical 7805 regulator are shown in Table 8.3. The meanings of the various specifications are as follows.

- **Potential:** This regulator uses and produces a positive voltage (some operate with negative voltages).
- **Case:** This tells us the type of case that is used as well as which terminal is which (see Figure 8.33b).
- V_{in} : The peak input voltage that is safe to use as a supply. Note that this can be considerably more than the output voltage, but the power associated with the drop of voltage will be wasted by the regulator and may cause it to overheat. Normally the input voltage should only be 2–3 V more than the desired output.
- V_{out} : The output voltage will be close to 5.0 V but may vary a little from one device to another. It will be steady and smooth at somewhere between 4.8 and 5.2 V for the 7805 regulator.
- I_{out} (nom): The regulator is designed to supply a ‘nominal’ current of 1 A. However, the maximum current that can be supplied may be more than this depending on how well the heat produced in the regulator is dissipated. If the regulator gets too hot, it automatically shuts down until it cools off. Heat sinks are often used to help disperse the heat produced.
- **Dropout voltage:** This is the minimum voltage drop between the input and the output. In order to produce the regulated 5 V output, the regulator needs at least 7 V input. Any ripple in the supply must not go lower than this.
- **Power dissipation:** The amount of power the component can handle depends on the ambient temperature and the heat sink used. With no heat sink, at normal temperatures, this 7805 regulator can handle about 3 W, but with a very good heat sink it can handle up to 20 W.
- **Maximum temperature:** Any electronic component will fail if it gets too hot. These regulators will automatically shut down to prevent damage if the power dissipated causes an excessive temperature rise.

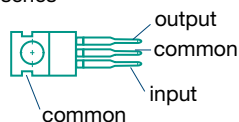
Table 8.3 The specifications for a typical 7805 voltage regulator

Potential	positive
Case	TO-220
V_{in}	max, 35 V
V_{out}	min, 4.8; typ, 5.0; max, 5.2
I_{out}	nominal 1 A
Dropout voltage	2 V
Power dissipation	3–20 W
Maximum temperature	125°C



(b)

TO-220 package
78XX
Positive regulator
series



TO-3 package
78XX
Positive regulator
series

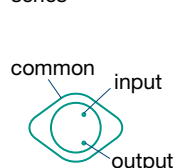
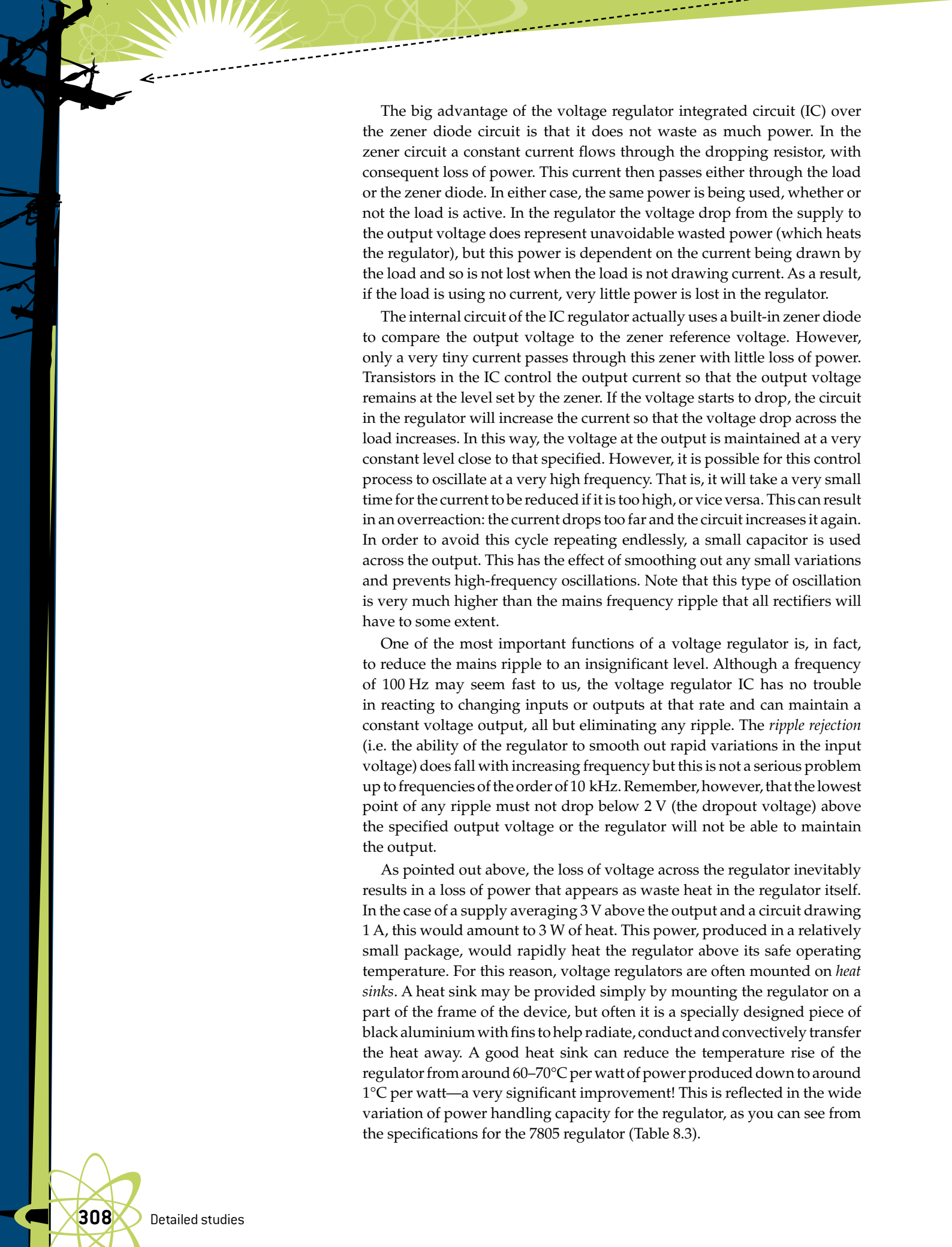


Figure 8.33 (a) These are typical voltage regulators which can be bought for around \$1 or \$2.
(b) Typical package and pin specifications.



The big advantage of the voltage regulator integrated circuit (IC) over the zener diode circuit is that it does not waste as much power. In the zener circuit a constant current flows through the dropping resistor, with consequent loss of power. This current then passes either through the load or the zener diode. In either case, the same power is being used, whether or not the load is active. In the regulator the voltage drop from the supply to the output voltage does represent unavoidable wasted power (which heats the regulator), but this power is dependent on the current being drawn by the load and so is not lost when the load is not drawing current. As a result, if the load is using no current, very little power is lost in the regulator.

The internal circuit of the IC regulator actually uses a built-in zener diode to compare the output voltage to the zener reference voltage. However, only a very tiny current passes through this zener with little loss of power. Transistors in the IC control the output current so that the output voltage remains at the level set by the zener. If the voltage starts to drop, the circuit in the regulator will increase the current so that the voltage drop across the load increases. In this way, the voltage at the output is maintained at a very constant level close to that specified. However, it is possible for this control process to oscillate at a very high frequency. That is, it will take a very small time for the current to be reduced if it is too high, or vice versa. This can result in an overreaction: the current drops too far and the circuit increases it again. In order to avoid this cycle repeating endlessly, a small capacitor is used across the output. This has the effect of smoothing out any small variations and prevents high-frequency oscillations. Note that this type of oscillation is very much higher than the mains frequency ripple that all rectifiers will have to some extent.

One of the most important functions of a voltage regulator is, in fact, to reduce the mains ripple to an insignificant level. Although a frequency of 100 Hz may seem fast to us, the voltage regulator IC has no trouble in reacting to changing inputs or outputs at that rate and can maintain a constant voltage output, all but eliminating any ripple. The *ripple rejection* (i.e. the ability of the regulator to smooth out rapid variations in the input voltage) does fall with increasing frequency but this is not a serious problem up to frequencies of the order of 10 kHz. Remember, however, that the lowest point of any ripple must not drop below 2 V (the dropout voltage) above the specified output voltage or the regulator will not be able to maintain the output.

As pointed out above, the loss of voltage across the regulator inevitably results in a loss of power that appears as waste heat in the regulator itself. In the case of a supply averaging 3 V above the output and a circuit drawing 1 A, this would amount to 3 W of heat. This power, produced in a relatively small package, would rapidly heat the regulator above its safe operating temperature. For this reason, voltage regulators are often mounted on *heat sinks*. A heat sink may be provided simply by mounting the regulator on a part of the frame of the device, but often it is a specially designed piece of black aluminium with fins to help radiate, conduct and convectively transfer the heat away. A good heat sink can reduce the temperature rise of the regulator from around 60–70°C per watt of power produced down to around 1°C per watt—a very significant improvement! This is reflected in the wide variation of power handling capacity for the regulator, as you can see from the specifications for the 7805 regulator (Table 8.3).

Worked example 8.3D

The need for a voltage regulator to be mounted on a heat sink depends very much on the conditions of its use.

- a Taking the specifications for the 7805 regulator given in Table 8.3, and assuming a maximum of 12 V input, what is the maximum current output that could be used with an unmounted 7805 regulator?
- b Given a very good heat sink, a 7805 can handle up to about 2 A of current. At that current, what is the maximum input voltage that can be used?

Solution

- a The maximum power dissipation of the unmounted regulator is 3 W. The power which needs to be dissipated is the difference between the input and output voltage:
 $\Delta V = 12 - 5 = 7 \text{ V}$
$$I = \frac{P}{\Delta V}$$
$$= \frac{3}{7} = 0.4 \text{ A}$$

Hence, the maximum current available is 0.4 A.
- b If the regulator is well mounted on a heat sink, it can dissipate 20 W. If a current of 2 A is flowing:
$$\Delta V = \frac{P}{I}$$
$$= \frac{20}{2} = 10 \text{ V}$$

Hence, the voltage *difference* is limited by the 20 W to 10 V. As the output voltage is 5 V, the maximum input voltage will be 15 V.

You may have noticed in Worked example 8.3D that if the input voltage was only, say, 7 V and the regulator was mounted on a good heat sink, theoretically, the maximum current could be up to 10 A (as $P = \Delta V \times I$ so $20 \text{ W} = 2 \text{ V} \times 10 \text{ A}$). This is not the case, however. Other considerations in the circuitry inside the regulator limit the current to a maximum of about 2.5 A. The added power-handling capacity when a heat sink is used enables higher input voltages to be used, but it does not allow much larger currents.

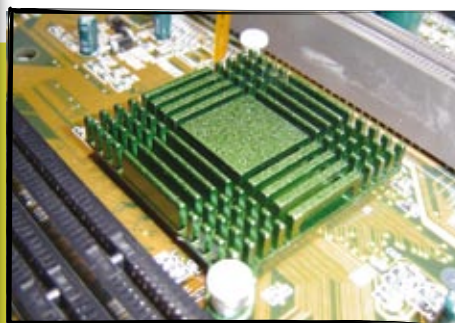
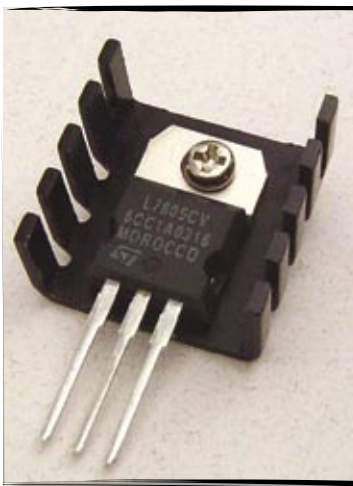


Figure 8.34 A heat sink can help dissipate the unavoidable wasted heat energy produced in integrated circuits. (a) A voltage regulator. (b) A computer CPU.



8.3 summary

Rectification and power supplies

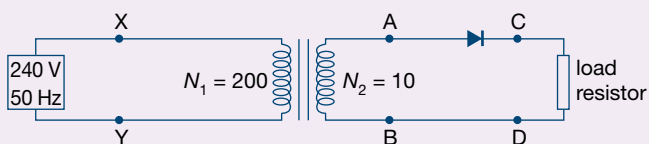
- Most electronic devices require a low-voltage DC to be obtained from the mains, high-voltage AC.
- The first step in this process is the use of a transformer to provide low-voltage AC.
- The low-voltage AC has to be rectified, i.e. converted to a half- or full-wave varying DC by the use of diodes which allow current to flow only in one direction.
- A four-diode bridge circuit is the most common way of producing a full-wave DC voltage.
- A large value capacitor is required to smooth out the full-wave DC.
- Even with a large capacitor, the output from a rectifier circuit may still have some 100 Hz ripple.
- A zener diode can be used to stabilise and remove the ripple from the voltage across a varying load.
- A voltage regulator IC is the best way to obtain a stabilised, steady, ripple-free DC voltage from a rectifier circuit.



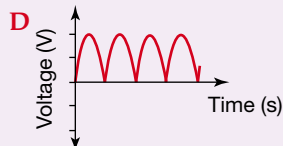
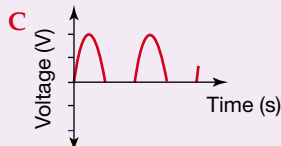
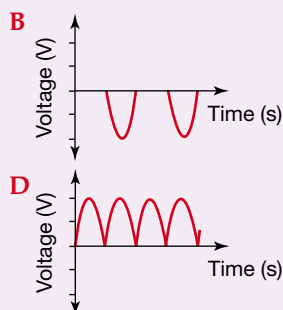
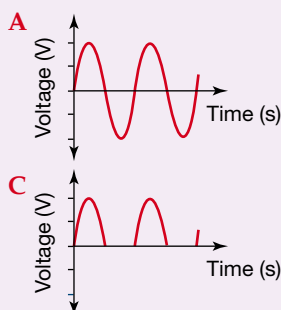
8.3 questions

Rectification and power supplies

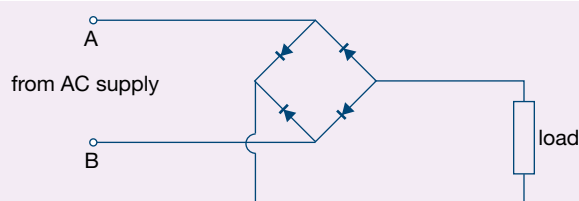
- 1 As part of a research project, a physics student designs and builds the following circuit.



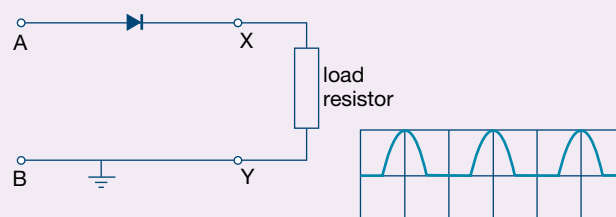
- What is the function of this circuit?
- What is such a circuit commonly called?
- The student examines the characteristics of the circuit using a CRO. The input to the primary winding of the transformer is the mains supply. Which of the following diagrams A–D best describes the display when the CRO is connected across:
 - XY?
 - AB?
 - CD?



- 2 A full-wave rectifier circuit is depicted in the following diagram.



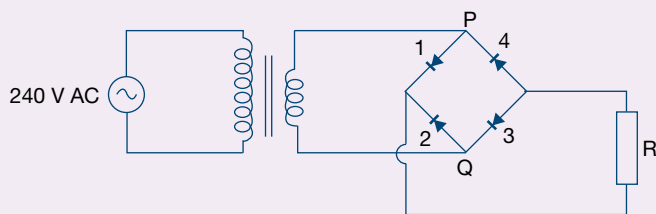
- Draw the path for the current through the bridge rectifier if A is the higher potential at the supply.
 - The AC now changes polarity, so that B is positive. Use a different colour to mark the path of the current path through the bridge rectifier now.
- 3 The following graph depicts the waveform seen when the input of a CRO is placed across the terminals XY of the circuit shown. The timebase setting of the CRO is 100 ms cm^{-1} and the grid is $1.0 \times 1.0 \text{ cm}$.



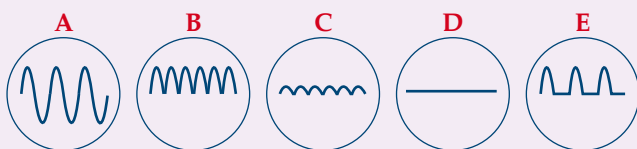
- During which time intervals does the CRO display show that the diode is forward biased?
- During which time intervals does the CRO display show that the diode is reverse biased?
- Draw the display that would be observed if the CRO input was connected across AB.
- What is the frequency of the input signal?



The following information applies to questions 4 and 5.
A CRO is used to investigate the voltage at various points in the following circuit.



- 4 Which of the following diagrams A–E best describes the resulting trace if the CRO probes are placed:
a on points P and Q?
b across the load resistor R?



- 5 A large capacitor is now placed across the load resistor R.
a With the CRO probes still across R, describe the changes you expect to see in the trace.
b If the capacitor is replaced by a smaller one, how would the trace change?
c Why is a capacitor used across the output of a rectifier circuit?
- 6 A 2000 μF capacitor is used across the output of a full-wave rectifier circuit producing a peak of 6 V from a 50 Hz AC input. The circuit is to be used to supply current to a load which will draw between 10 mA and 1 A.
a Approximately how much ripple would you expect to see in the output at the minimum and maximum load currents?
b How much charge is stored in a 2000 μF capacitor at 6 V? For how long would this charge last if it were to provide a constant current of 10 mA, and of 1 A?
c Considering your answer to part b, why will there be any ripple voltage at all?
d What could be done to reduce the amount of ripple in the output voltage?
- 7 A 12 V car battery is to be used to run a 9 V radio by using a dropping resistor and zener diode in a circuit similar to that shown in Figure 8.31. The radio may use between 50 and 200 mA of current.

- a** What should be the resistance of the dropping resistor?
b If the radio is using only 50 mA of current, how much power is it using, and how much power is being wasted in this circuit?
c Would the radio operate correctly if a lower or higher value for the dropping resistor was used? Illustrate your answer with some numerical examples.

- 8 What are the disadvantages of a zener diode-based circuit when used to stabilise the voltage across a load which draws a varying current?
- 9 Which (one or more) of the following characteristics are true of a voltage regulator such as the 7805 discussed in the text?
A A voltage regulator can be used instead of a capacitor to smooth the output of a full-wave rectifier.
B The use of a voltage regulator avoids the need for a very large capacitor on the output of a full-wave rectifier.
C A voltage regulator does not waste any power from the output of a full-wave rectifier.
D A voltage regulator uses very little power when the load circuit is not drawing current.
E A voltage regulator will be damaged if the load circuit draws too much current.
- 10 These questions relate to a 7805 voltage regulator. See Table 8.3 for the specifications of this regulator and Figure 8.32 for a diagram of a circuit containing a regulator.
a Why is a 2500 μF capacitor shown in the circuit diagram?
b Given that the output is 5 V, why is 8.5 V needed from the transformer?
c What are the limiting factors on the amount of current that can be drawn from this circuit?
d Why would it not be advisable to use a transformer which gave a peak output of 25 V in this circuit?
e How much power would be wasted as heat if the input to the 7805 was 12 V and the current drawn was 500 mA?

8.4

Constructing and testing a working power supply

Now that we have all the elements of our power supply, it is time to put them together and test the results! There are a number of ways in which circuits can actually be constructed. Perhaps the cheapest and simplest is to solder the components directly onto tag strips available from electronics shops. Zinc-plated nails in a piece of chipboard can serve the same purpose. A more satisfactory and flexible approach is to use electronic breadboards into which the leads of the components can be inserted without any need for solder. These consist of an array of small holes that are electrically connected in rows. Components to be joined are simply inserted into the same row. More permanent arrangements can be made with various types of 'experimenter's board' or 'stripboard' which has pre-drilled holes and copper strips on one side. The copper strips are used to join the components but can be cut with a knife or drill to isolate different parts of the circuit.

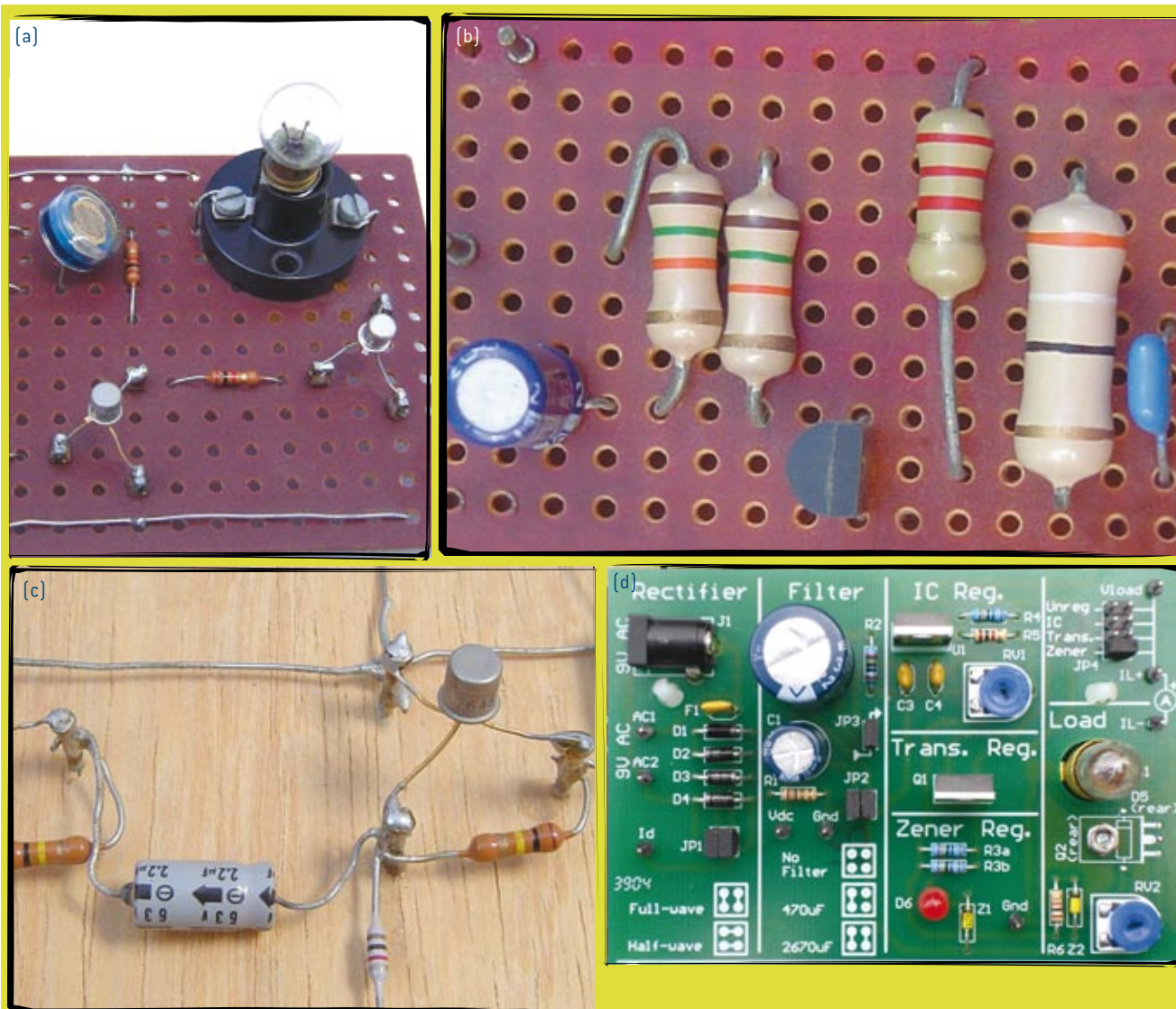



Figure 8.35 Different ways of constructing an electronic circuit. (a) Matrix board and pins. (b) Stripboard with copper strips underneath. (c) Plated nails and wood. (d) A breadboard makes experimenting with circuits easy. The holes in the short rows are electrically connected.



To save time, there are ready-made kits which only require connections with alligator leads or similar. Even further removed from the real thing are the 'virtual circuits' that can be built on a computer screen. While this project could be done with kits or computers, these approaches do not allow the possibility of dry joints, failed components, burn-outs and other human errors which make tinkering with electronics so much fun! More seriously, part of the purpose of this project is to experience the sorts of problems that afflict real-world circuits and to develop the skills that are needed to troubleshoot and solve problems. For these reasons we suggest using either the breadboard or the stripboard approach. The kits and computers may, however, complement the real thing by providing the opportunity to see what happens with different components; it is cheaper (and safer) to overload components in a virtual circuit!

While this section contains fairly specific instructions on building the power supply, it would be in the spirit of this study to try to think ahead, and design and build your own circuits based on the ideas presented in the previous sections. Provided you take reasonable precautions, the worst you can do is burn out a few components! Read the following Physics in action carefully before starting any experimental work.

Physics in action

Watch for hot spots!

When building any electronic circuit, there are a few basic precautions you need to take to avoid potential problems.

- Only use low-voltage AC or DC supplies (up to 25 V) or batteries to power your project.
- Design your circuit on paper first and check it carefully or have it checked by your teacher or other competent person.
- Check the components you are using to ensure that they are what you think they are and that they are not burnt out, open circuit or short circuit. Resistors should be checked with a multimeter. Capacitors should also be checked for short circuits with a multimeter; large capacitors may show a significant current flow for a few moments as they charge up from the battery in the multimeter. **WARNING: Make sure capacitors are discharged before checking with a multimeter or you may damage the meter!** Diodes should be checked to ensure that their resistance is very low one way around but very high the other way.
- Make sure that any electrolytic capacitors are put in the circuit the right way around. They normally have an arrow that points to the negative end. This end should be connected so that it is closer (in voltage, that is) to the zero, or ground, line. An electrolytic capacitor can actually explode if connected the wrong way around!
- If you are using a soldering iron, be very aware of where you put the hot tip when you put it down. It can give you a nasty burn. Be particularly careful that the tip is nowhere near its own power lead; it could easily burn through the insulation and cause a short circuit or expose a live wire.
- When soldering small components, they will get very hot. It is important not to overheat them so be quick and definite with the soldering iron, but make sure the solder flows around the joint well. When soldering semiconductors, it is best to hold the lead being soldered with a small pair of pliers or a special heat-sink clip to avoid too much heat reaching the crucial parts of the device.
- Before connecting any power to the circuit you have built, double check it against your circuit diagram.
- When you turn on the power, the first principle in checking any electronic circuit is to watch for hot spots. Excessive heat is a sure sign of a fault somewhere. If you see or smell smoke, or feel a hot component, turn off the AC supply straight away and check the circuit. Remember that components such as diodes and voltage regulators will normally be warm to hot, but not excessively so. Check that the diodes in a bridge circuit are all equally warm. If one is hotter than the others, there may be a problem.
- If possible, arrange to have an ammeter measuring the current from the power supply and a voltmeter measuring its voltage. If, when you first power up the circuit, there is any marked deviation from what you expect for the current and voltage from the power supply, turn off the circuit immediately and check for possible wiring errors.

Table 8.4 Parts for a simple regulated power supply

Mounting material	Tag strips, breadboard, stripboard or other means
Source of low-voltage AC	Mains-operated transformer or school power supply with AC outlet (usually the yellow terminals). Output in the range 8–12 V
Diodes	There are several suitable types of diodes. A common one is the 1N4004 which will handle up to 1 A. You may want one that has a higher current rating, such as the 1N5404 (3 A).
Rectifier bridge	This is an alternative to individual diodes. A W04 (1 A) or P04 (6 A) or other medium power type is suitable.
Capacitor (electrolytic)	A range of values from about 100 μF to 2200 μF would be suitable so that you can experiment to see the effect of different values. A working voltage rating of 16 V or 25 V is satisfactory unless your power supply is designed to give higher voltages.
Voltage regulator	A 78XX voltage regulator (where XX is the voltage required, e.g. a 7805 for a 5 V supply). The TO-220 package type is cheaper and easier to mount on a heat sink.
Heat sink	Depending on the current you expect to use, you may need a heat sink for the regulator. This could be a simple piece of aluminium or a purpose-built component with fins to radiate the heat away. Alternatively, it might be possible to bolt the regulator to an aluminium case.
Capacitors (in/out)	0.1 μF (or thereabouts) ceramic or tantalum. Again, different values could be tried.
Load	Many things could be used as a load: light bulbs, a motor, a radio, or just resistors. A wire-wound variable resistor (say 500 or 1000 Ω , 3 W) would provide a good variable load for your experiments. A cheaper alternative would be a range of normal resistors—but consider the power dissipation required at lower resistance values where the currents will be higher.

The basic rectifier

The first task is to build a basic full-wave rectifier circuit. As we have described in section 8.3, this can be based on a low-voltage (say 8–12 V) AC supply from a power pack or a commercial transformer. (Do not attempt to wire the primary of a transformer to a mains fitting!) We then simply need to construct the four-diode bridge circuit as described in the last section and shown in the circuit diagram in Figure 8.36. Figure 8.37 shows one way in which this diode bridge can be constructed on a breadboard, but before you look at that photograph, perhaps you could try to work it out for yourself; there are other possibilities.

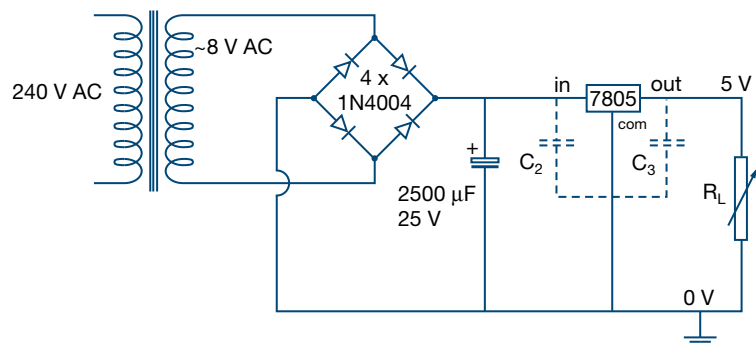


Figure 8.36 The circuit diagram for our complete regulated power supply circuit. The various components are described in the text.

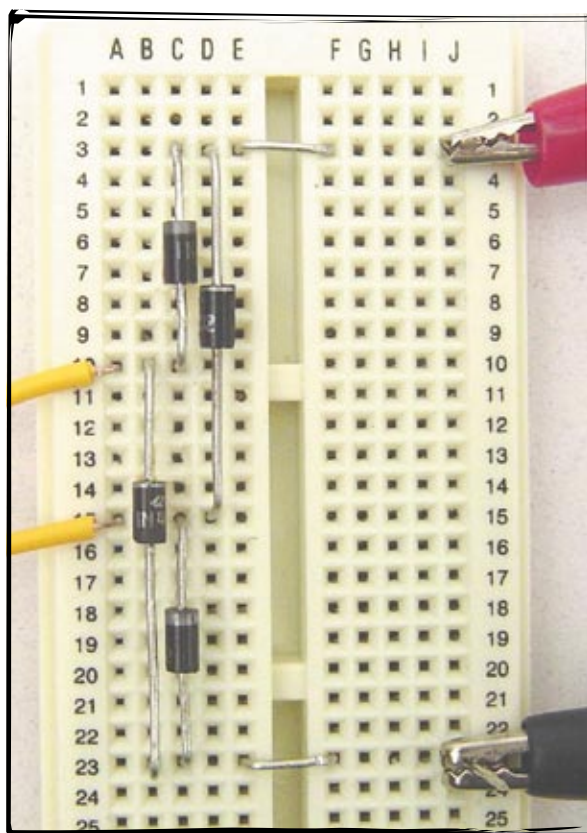


Figure 8.37 One way to construct the bridge rectifier circuit on breadboard. The short rows of five holes are all connected under the surface.

Before proceeding, confirm that the bridge circuit is doing what we expect. The circuit will not operate correctly unless there is a load resistor in place. A voltmeter placed across the output may provide a suitable load, but it is better to fix a temporary load resistor into place. This should be a fairly high value at this point ($>1\text{ k}\Omega$ for example). When turned on, there should be a DC voltage across the load. If you have access to a CRO, look for the characteristic full-wave rectifier pattern (as shown in Figure 8.25).

Smoothing it out

Once we have confirmed that the diode bridge is functioning properly, it is time to smooth out the full-wave voltage pattern by adding the capacitor. A large electrolytic capacitor is required for this purpose. The value of the capacitor will depend on the current we expect to draw and the difference between the regulator voltage and the peak output of the rectifier. While it is possible simply to use a high-value capacitor, these are expensive and bulky, so it is good practice to use the smallest value that is satisfactory. We can calculate the minimum value necessary by estimating the time constant required for the circuit, taking into account the expected load current and the 100 Hz input waveform. Alternatively (or, preferably, as well), experiment with different values once the circuit is complete.

Remember that the amount of ripple in the output will depend on the value of the load resistance, so you may wish to experiment with different load resistors at this stage. However, be careful not to overload the diodes in the bridge circuit. Check their maximum current rating and don't allow the load to draw more current than that. For example, 1N4004 diodes have a maximum current rating of 1 A, so the load resistance (in ohms) must not

Physics file

The rated transformer voltage will be the RMS (root mean square) voltage (see section 8.1). However, the capacitor will charge up to the peak voltage which is $\sqrt{2}$ times as much. There will be a small drop (about 1 V) across the two diodes in the circuit (for each half cycle) but the peak voltage delivered to the capacitor will still be quite a bit higher than the RMS value. Due to the ripple, the average value across the capacitor will be a little less than this, but for an 8 V transformer we could expect around $8 \times \sqrt{2} - 1 \approx 10\text{ V}$ on the capacitor.

be less than the output voltage (in volts) $V = IR$ where $I = 1$ A. If possible, include an ammeter in the load circuit to check the current flowing.

The voltage at the capacitor may be more than expected. If, say, we used an 8 V transformer, the voltage across the capacitor may be about 10 V. Can you see why this is? (Check with the Physics file on page 315.)

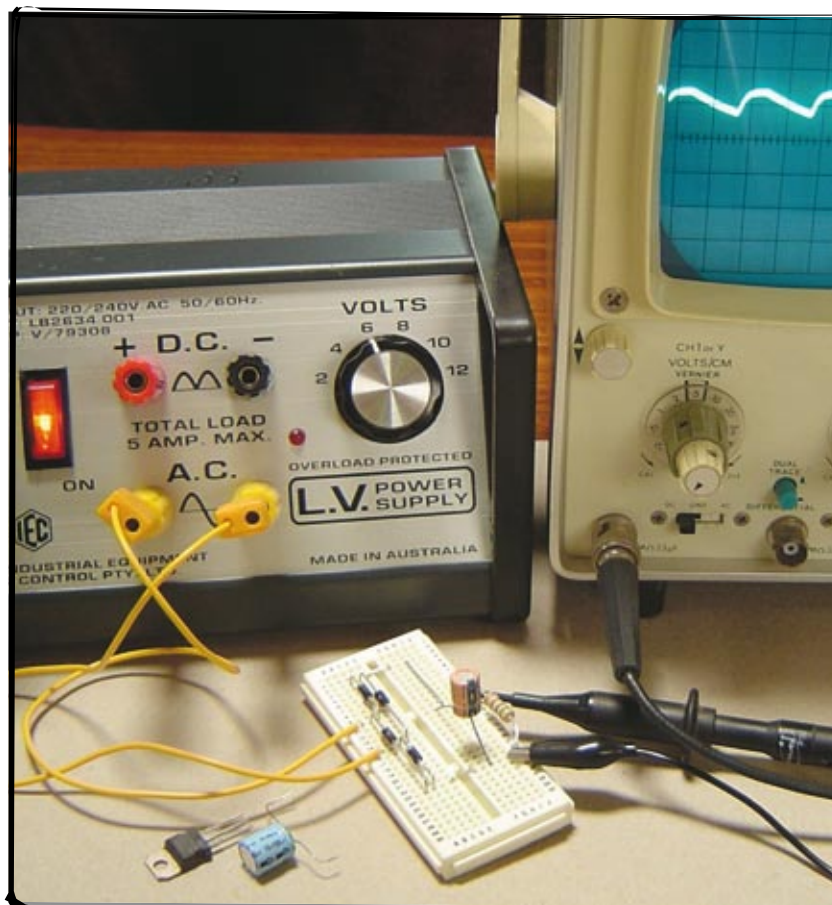


Figure 8.38 The smoothing capacitor has been added to the circuit. Note the polarity arrow (blue strip on this capacitor) must be toward the negative side of the circuit.

Now to regulate it

The output of the rectifier, smoothed by the capacitor, is now taken to the V_{in} terminal of the regulator (see Figure 8.29, page 305). The 'common' terminal is connected to the zero (ground) line and the output terminal to a suitable connection for the load. At this point it is necessary to decide whether to use a heat sink. In Worked example 8.3D, we saw how to work out the power which needs to be dissipated under various conditions of input voltage and output current. If the maximum power generated in the regulator will always be under about 3 W, then a heat sink will not be needed. For a practical power supply, however, it is normally advisable to use a heat sink of some sort.

To test the characteristics of this circuit, we now need to use either a variable resistor as the load (ensure that it can handle the power) or a set of discrete resistors of appropriate ratings. Start with a high resistance load and check to make sure the circuit is behaving as it should. Then gradually

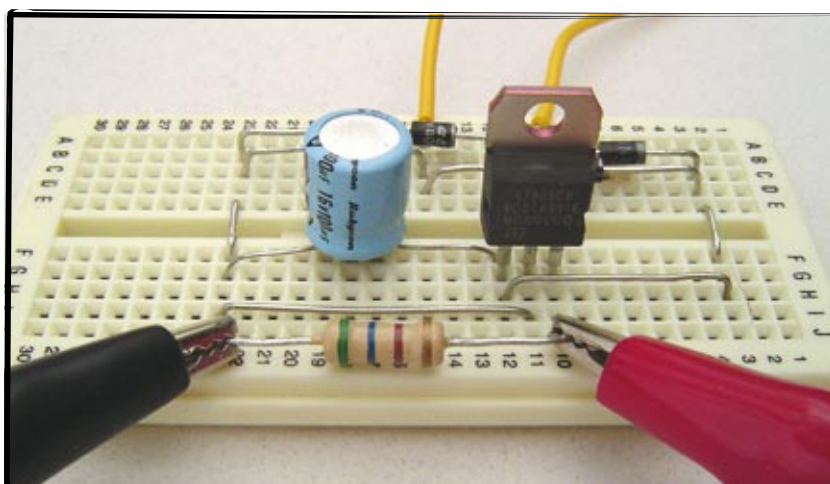


Figure 8.39 The 7805 voltage regulator has been added to the circuit. The resistor in the foreground is acting as a load.

reduce the load resistance, preferably measuring the output current and voltage at the same time to ensure that they are within the appropriate limits. Also, remember that the limit on the current available may well be the current rating for the diodes you have used. Check to make sure that they are not getting too hot. If possible, use a CRO to look at the output voltage to check for any ripple. If you can increase the gain of the CRO sufficiently (and use the AC input) you may detect some ripple.

You may have noticed that the circuit diagram (Figure 8.36) includes two low-value capacitors, one at the input to the regulator and one at the output. Provided the large electrolytic smoothing capacitor is physically close to the regulator, C_2 is not needed. Its purpose is to damp down any possible high-frequency oscillations in the input circuit of the regulator. Likewise, the C_3 capacitor will not usually be necessary, but is often used because these oscillations can be started by unpredictable circumstances in the circuit in which the power supply is operating (see the adjacent Physics file).

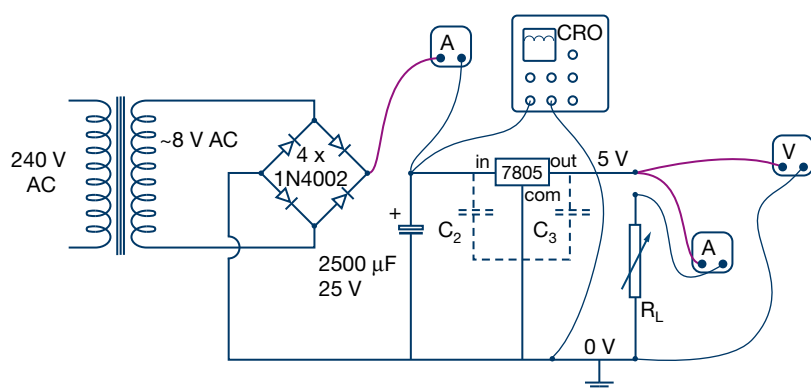
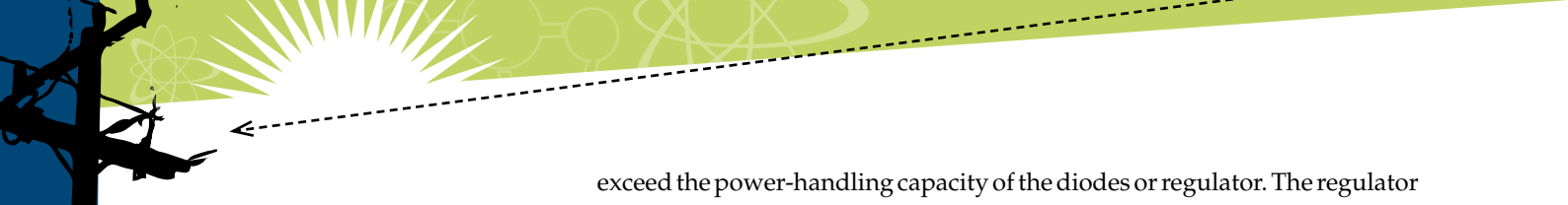


Figure 8.40 A set-up for testing a power supply. Changing the load will reveal the relationship between the output voltage and current. Changing the input AC voltage from the transformer, if possible, will reveal whether the regulator is doing its job.

If you have access to a variable AC supply (such as the typical school unregulated power supplies) you can check the function of your regulated power supply under conditions of varying input voltage. Be careful not to

Physics file

Capacitors C_2 and C_3 help to avoid oscillations. Any system that reacts to changes in the input or output is in danger of *positive feedback*. An example is the howl you hear if a microphone is placed near a loudspeaker. This can happen in the regulator, which reacts very quickly to any fall in the output voltage by increasing the current, and hence the voltage. If the voltage increases too much, the circuit will reduce it again. Under some conditions, usually only associated with high-frequency signals in the circuit, this process can get out of control and swing back and forth. The purpose of C_2 and C_3 is to short circuit any high-frequency signals before they can cause this kind of oscillation.



exceed the power-handling capacity of the diodes or regulator. The regulator will shut down if it gets too hot, but the diodes will simply burn out!

Fault finding

It is fairly normal that when a circuit is first built it does not work. If you have built the circuit as we have been discussing it in this chapter and it worked, congratulations! Some fault-finding techniques have already been noted at various stages: watch the supply voltage for signs of short circuits, look for hot spots, use a CRO to look for ripple, and so on.

The most likely cause of incorrect operation is faulty wiring or a faulty connection, so if it doesn't work, check the wiring carefully against the circuit diagram, and check that the diagram is correct. If you are not sure about a connection, check that there is no voltage between the two wires involved. Before using an ohmmeter to check that there is no resistance in the connection, make sure that the circuit is turned off and that the section is isolated, otherwise the meter could be damaged by currents from the circuit. Occasionally you will find that a component is faulty. One reason for this is that it was faulty when you started. Another reason, however, is that the circuit damaged it! Before replacing it, make sure that it is the appropriate value or type, and that it is placed correctly in the circuit; semiconductors and capacitors can be destroyed by mixing up the leads, for example.

Soldering is a skill that requires practice. One cause of faults in soldered circuits is *dry joints*. This refers to the lack of wetting by the solder that sometimes occurs if the joint was not hot enough. The components can appear joined, but there may be no electrical connection. These joints don't look correct under close examination and will usually break if pulled, and so this is the easiest way to check for dry joints. To avoid dry joints don't be tentative with the soldering iron: place it firmly on the joint, watch until the solder runs and wets the contacts, and then remove it.

The values and components in the circuit we have described are not particularly critical, but different combinations may need to be matched. If you are using a 16 V AC input, don't use an electrolytic capacitor rated at only 16 V, for example. If you are using a high-current voltage regulator, it is necessary to use high-current diodes (or diode bridge) as well.

Hopefully you have now successfully built a power supply that has worked! Although our circuit was designed for a particular purpose, the same basic principles are used in an enormous array of electronic devices. If you used a breadboard, you may want to make a more permanent power supply for yourself. You could even add meters and switches and mount it all in a nice case. It is not difficult to modify our fixed output voltage circuit so that it will give a variable (but constant) DC voltage output. Any good hobby electronics book will have the details. If you have understood the principles involved in building this system, you will find that you are well on the way to a good understanding of electronics. Not only can it be an enjoyable hobby, it can also lead to many different career paths.



8.4 summary

Constructing and testing a working power supply

- Various techniques can be used to construct a circuit. It is important to check immediately that no problems are occurring when first turning on a new circuit.
- A diode bridge circuit produces a full-wave output.
- A large capacitor smooths the full wave to produce a near smooth output, but with some ripple.
- A voltage rectifier is used to reduce the output from the capacitor, leaving a constant, ripple-free DC output at a voltage set by the voltage regulator.
- Building a circuit is something of an art. Practice is needed and clear thinking is required to find any faults. Good luck!



8.4 questions

Constructing and testing a working power supply

- 1 When first turning on a circuit you have just constructed, which one or more of the following actions should you take to check that there are no damaging faults present?

- A** Watch for any signs of smoke or hot spots.
- B** Measure the current being drawn from the battery or power supply.
- C** Watch for an excessive drop in the voltage of the power supply.
- D** Check the output of the circuit to see if it is operating as it should.

- 2 When Lachlan turned on his newly wired-up rectifier circuit, he found that one of the four diodes in the bridge rectifier was significantly hotter than the other three, even when there was no load connected.

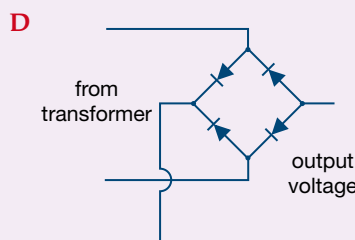
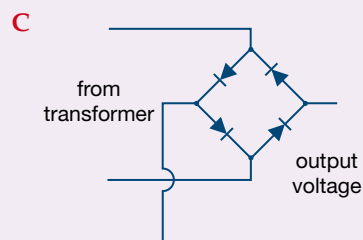
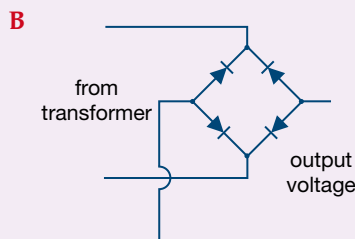
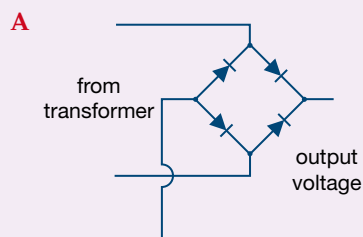
- a** How could a diode in the rectifier circuit become hot even when no current is flowing in the load?
- b** When a load is connected, Lachlan finds that little current is flowing. Can you explain why this is the case?

- 3 Does it matter which way around a capacitor is put into a circuit? Explain your answer.

- 4 A power supply circuit has been designed to produce a current of up to 1.5 A at a voltage which must remain within 10% of 5.0 V. It is to be tested using various values of wire-wound resistors as loads.

- a** When tested with a 100 Ω resistor, what should be the power rating of the resistor?
- b** What is the lowest load resistance that should be used, and what should its power rating be?
- c** Would it be safe to use a load resistor rated as 10 Ω , 1 W? Explain.

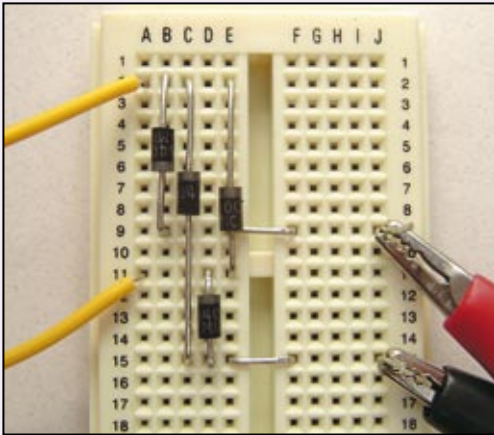
- 5 Diagrams A–D represent four different attempts to connect up a full-wave bridge rectifier circuit. (There may be one or more answers, or none, to each of the following questions.)



- a** Which of these circuits will produce a full-wave DC output?
- b** Which, if any, of these circuits is likely to result in damage to the diodes or transformer?
- c** In which, if any, of the circuits will no current flow from the transformer?



- 6 The photograph shows an attempt to connect up a full-wave bridge rectifier using a breadboard.



The holes in each of the short rows of five are connected under the surface of the breadboard. The two yellow wires are from the transformer and the red and black wires are the + and – DC output connections respectively. There is one wiring error in this circuit.

- Where is the error?
 - What would happen if this circuit was turned on?
 - How should the error be corrected?
- 7 A rectifier is to be used to supply a $1\text{ k}\Omega$ load resistor with a constant 5 V DC .
- Which of the following capacitors connected across the output would give the smoothest DC supply?
A $100\text{ }\mu\text{F}$
B $1000\text{ }\mu\text{F}$
C $10000\text{ }\mu\text{F}$
 - Which of these capacitors is an engineer most likely to use in a practical circuit?
A $100\text{ }\mu\text{F}$
B $1000\text{ }\mu\text{F}$
C $10000\text{ }\mu\text{F}$
 - Explain your answer to part b.
- 8 A transformer which supplies 5 V RMS is used to power a bridge rectifier circuit with a smoothing capacitor. What would be a reasonable estimate (to one significant figure) of the likely DC output voltage under low-load current conditions? Explain your answer.
- 9 Explain why the use of a voltage regulator in a rectifier circuit enables the use of a smaller smoothing capacitor.
- 10 A voltage regulator inevitably wastes a certain amount of power.
- Explain why this is the case.
 - What happens to this power and why is it a problem?
 - What is done to avoid this problem?

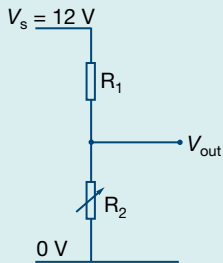


chapter review

Multiple-choice questions

The following information applies to questions 1–3.

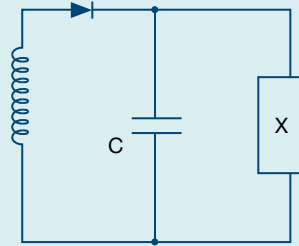
Aaron builds a voltage divider circuit which consists of one fixed resistor [R_1] and one variable resistor [R_2] placed across a 12 V supply, as shown in the diagram. He then tests the circuit by connecting a voltmeter between the output terminal and the zero voltage line.



- When he first measures the voltage at the output, he finds that it is equal to the 12 V supply voltage. Which one or more of the following could account for this observation?
 - R_1 has burnt out and the circuit is open.
 - R_2 is damaged and the sliding contact is not contacting the resistor at all.
 - R_2 is set to zero resistance.
 - R_2 is set so that it is equal to R_1 .
- He then rearranges the circuit and finds that the output voltage is 6 V. Which one or more of the alternatives above could account for this observation?
- He carefully sets $R_1 = 400\text{ k}\Omega$ and $R_2 = 100\text{ k}\Omega$ and then finds that the reading on his analogue voltmeter is 2.1 V, which is less than he expected. Which of the following options gives the reading he expected as well as a feasible explanation of why the actual reading could have been less?
 - Expected 9.6 V. The resistance of the meter was significant compared to R_1 and R_2 .
 - Expected 3.0 V. The resistance of the meter was significant compared to R_1 and R_2 .
 - Expected 2.4 V. The resistance of the meter was significant compared to R_1 and R_2 .
 - Expected 2.4 V. The meter added resistance to the circuit and so R_2 was effectively greater than $100\text{ k}\Omega$.

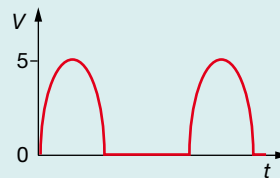
The following information applies to questions 4–7.

A simple rectifier circuit, supplied from a transformer connected to the 50 Hz mains, and as shown in the diagram, is set up to supply a DC current to a small electronic circuit shown as X. The capacitor C can be connected in the circuit as shown. The circuit X draws a current from zero up to a maximum of 10 mA. It requires a DC voltage of 5 V with ripple less than 10% to operate correctly.

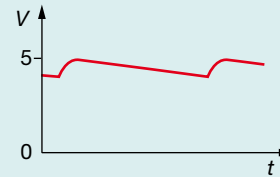


- Which of the following graphs (A–D) would best represent the voltage across X without the capacitor connected?

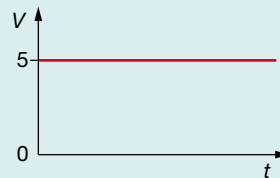
A



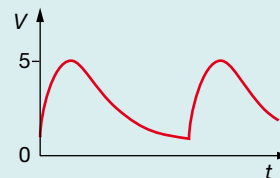
B



C



D



A $30\text{ }\mu\text{F}$ capacitor is now connected in the position indicated.

- Which of the graphs (A–D above) would best represent the voltage across X while X draws a very small current of 0.1 mA?
- Which of the graphs (A–D above) would best represent the voltage across X while it draws its maximum current of 10 mA?

7 Which of the following best explains whether the circuit will operate correctly with the $30\ \mu\text{F}$ capacitor in place?

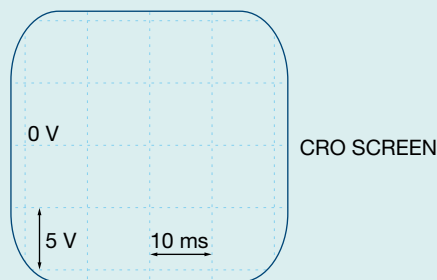
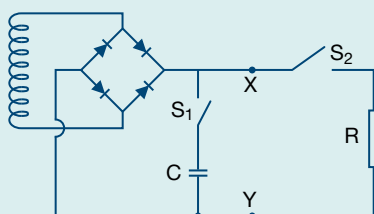
- A It will operate satisfactorily as the ripple will always be less than 10%.
- B It will operate satisfactorily as the ripple will be less than 10% most of the time.
- C It will not operate satisfactorily as the ripple will be more than 10% at times.
- D It will not operate satisfactorily as the ripple will always be more than 10%.

8 The symbol for a rectifier diode is in the shape of an arrow. The arrow represents the direction:

- A in which the conventional current will flow
- B in which the electrons flow through the diode
- C in which it should be connected in the circuit
- D of none of the above.

The following information applies to questions 9–15.

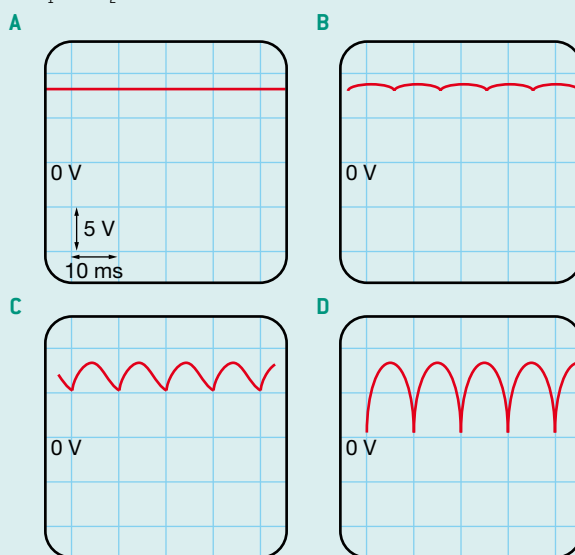
A full-wave bridge rectifier circuit is constructed as shown in the following circuit.



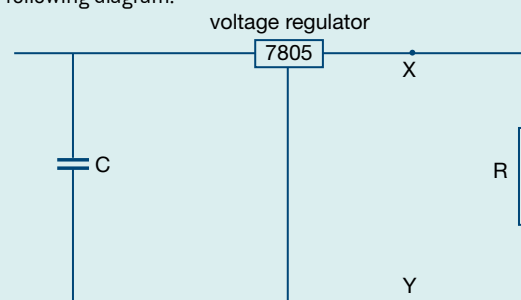
S_1 connects C, a $200\ \mu\text{F}$ capacitor, into the circuit, and S_2 connects R, a $500\ \Omega$ resistor, into the circuit. The AC output of the transformer is $6.0\ \text{V RMS}$ at $50\ \text{Hz}$. A CRO is used to look at the waveform between points X and Y. Which of the graphs A–D below best represents the waveform you would see with:

- 9 S_1 and S_2 both open
- 10 S_1 closed but S_2 open

11 S_1 and S_2 both closed.



In order to produce a more satisfactory DC output from the circuit in the previous questions, a $5\ \text{V}$ voltage regulator is added, as shown in the following diagram.



12 Which of the following best describes the waveform seen on the CRO connected to X and Y now?

- A A straight line at about $5\ \text{V}$
- B A straight line at about $6\ \text{V}$
- C A straight line at about $7.5\ \text{V}$
- D A straight line at about $8.5\ \text{V}$

13 Which of the following is the best estimate of the current flowing through the $500\ \Omega$ load resistor?

- A A DC current of less than $5\ \text{mA}$
- B A DC current of between 5 and $20\ \text{mA}$
- C A DC current of $20\ \text{mA}$ or more
- D An AC current

14 The power being dissipated in the voltage regulator would be:

- A negligible
- B less than $40\ \text{mW}$
- C more than or about $40\ \text{mW}$
- D enough to shut down the regulator unless it has a very large heat sink.

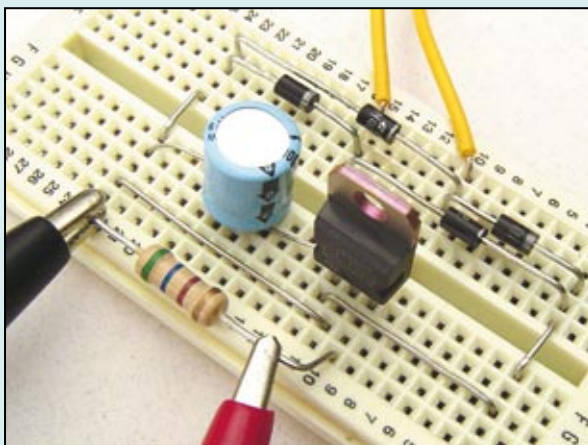
15 For this voltage-regulated power supply, which one or more of the following will result in a change in the current flowing through the load resistor?

- A The voltage of the supply to the transformer is increased.
- B The value of the capacitor used is increased.
- C The value of the load resistance is increased.
- D The diodes are changed to ones with a lower conduction voltage.

16 Two power supply circuits are built. Circuit X is constructed using a full-wave diode bridge and a very large value smoothing capacitor so that no significant ripple is present in the output voltage over the range of currents required. Another similar circuit, Y, is built using a smaller capacitor but includes a voltage regulator. In comparing these two circuits, which **one or more** of the following will be true of circuit X in comparison to circuit Y?

Circuit X will:

- A be physically smaller and less expensive
 - B waste less energy as heat will not be lost in the regulator
 - C be more stable against voltage changes in the transformer
 - D produce a higher output voltage from the same transformer and rectifier.
- 17 The photograph shows an attempt to connect up a full-wave bridge rectifier using an electronic breadboard. The holes in each of the short rows of five are connected under the surface of the breadboard. The two yellow wires are from the transformer and the red and black clips are the + and – DC output connections, respectively. There is one wiring error in this circuit.



Which one of the following best describes the wiring error in this circuit?

- A One of the diodes is the wrong way around.
- B There is a missing connection from the regulator to the + output.
- C The load resistor is in the wrong place.
- D The capacitor is connected the wrong way around.

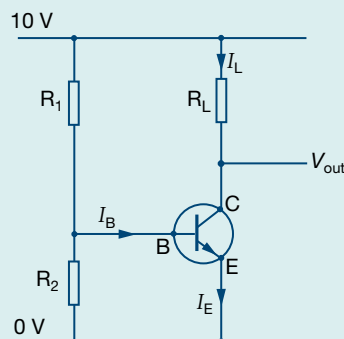
The following information applies to questions 18–20.

For each of the possible faults (questions 18–20) in a voltage-regulated power supply circuit, choose **one or more** of the following possible effects.

- A Half of the voltage peaks would disappear.
 - B The diodes could be overloaded and burn out.
 - C The output voltage would drop.
 - D Ripple will appear at the output.
- 18 The smoothing capacitor becomes open circuit (with no capacitance).
- 19 The smoothing capacitor breaks down and becomes a short circuit.
- 20 One of the diodes burns out and becomes an open circuit.

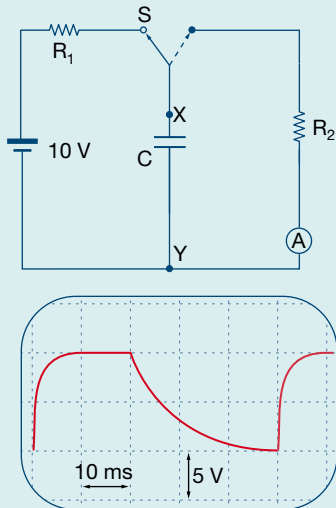
Extended-answer questions

21 The circuit in the following diagram is constructed using three resistors (R_1 , R_2 and R_L) and a transistor connected to a supply voltage of 10.0 V. The transistor has three connections labelled C, B and E. Knowledge of the way the transistor works is not required for this question. When R_1 is set to 1.9 k Ω , the current through resistor R_1 is found to be 5.0 mA. The current through R_2 is found to be 4.0 mA and the output voltage $V_{out} = 4.0$ V.



- a What is the value of R_2 ?
- b A current of 51 mA is found at the E terminal of the transistor (I_E). What is the current in R_L ?
- c What is the value of the load resistor, R_L ?

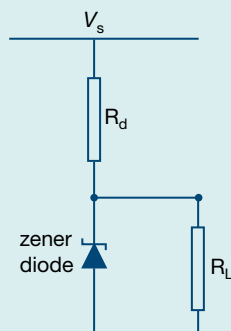
- 22 The following circuit is designed to repeatedly charge and discharge a capacitor C . S is a switch that oscillates rapidly back and forth. The capacitor is charged through R_1 from a 10 V supply and then discharged through R_2 and an ammeter. The time for the whole cycle is 50 ms. A CRO is connected across the capacitor at points X and Y and the trace obtained is shown in the following diagram. It is found that the average current recorded by the ammeter is 4 mA. Assume that the power supply and ammeter are 'ideal'.



- How many discharge cycles are there each second?
- What is the charge that passes through the ammeter each cycle?
- What is the value of the capacitance of the capacitor?
- Estimate the time constant for the discharge cycle.
- Hence, estimate the value of the resistor R_2 .
- Which of the following would be the best estimate for the value of R_1 : 10, 100, 500 or 1000 Ω ? Explain your choice.

- 23 A 22 k Ω resistor, known to be accurate to 1%, is placed across a regulated power supply of 10.0 V. A multimeter is then used to measure the current flowing and is found to give a reading of 0.43 mA. Assuming the reading to be accurate, what can you say about the multimeter? Include appropriate calculations in your answer.

- 24 A zener diode with a breakdown voltage of 8.0 V is used in the following circuit to ensure a constant voltage of 8.0 V across the load resistor R_L . R_d is a dropping resistor used to reduce the supply voltage V_s down to an appropriate level. The value of the load resistance is initially equivalent to 400 Ω and the supply voltage is 12 V.

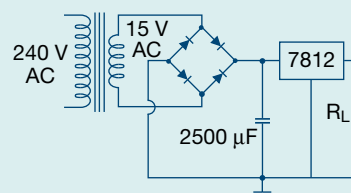


- Under these conditions, what would be the best value for R_d and how much current would flow through the zener diode?
- Given this value for R_d , if the supply voltage then increases to 15 V, how much current will flow in the zener diode?
- With the supply voltage restored to 12 V, and the same R_d , the value of the load resistance increases to 1000 Ω . What will be the current in the zener now?

- 25 Some faults in diodes or capacitors can result in them becoming a short circuit. This could lead to damage in other parts of the circuit and other serious consequences. Describe some of the possible consequences and some ways in which they can be avoided in practical power supply circuits.

- 26 The specifications of a 7805 voltage regulator say that the output voltage will typically be 5.0 V, with a minimum of 4.8 V and a maximum of 5.2 V. They also state that the line regulation is typically 3 mV provided $7 < V_{in} < 25$ and other conditions are fulfilled. Explain what is meant by each of these figures. Does this mean that the output voltage of a regulator in a circuit can vary between 4.8 and 5.2 V?

- 27 The following circuit represents a typical power supply circuit designed to produce a 12 V DC output across a resistive load R_L at currents up to 1.5 A. The transformer produces a 15 V RMS AC output. The diodes conduct with a forward voltage of 0.5 V across them in the range of current required. The 7812 regulator requires a minimum of 14.6 V to produce a 12 V output and has a maximum input voltage rating of 27 V. With no heat sink it can dissipate about 6 W, but with a very good heat sink it can dissipate about 20 W.



- What is the peak output voltage of the transformer?
- To what voltage will the capacitor charge if little current is being used by the load?
- Estimate the amount of ripple in the input to the regulator if the full 1.5 A is being drawn by the load.
- Will the regulator operate satisfactorily and be able to maintain a smooth 12 V output?
- How much power is being dissipated in the various parts of this circuit at full current output?
- Will the 7812 regulator require a heat sink in this circuit?

Unit

area of study 1

Unit

Electric power

outcome

On completion of this area of study, you should be able to explain the operation of electric motors, generators and alternators and the generation, transmission, distribution and use of electric power.

Magnets and electricity

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 area of study is an essential first step in meeting that challenge.

This photo shows an aurora, one of nature's most beautiful phenomena. They are produced when electrically charged particles from the Sun enter the Earth's magnetic field. The photos illustrate two different aspects of the essential importance of electromagnetism in our world.

by the end of this chapter

you will have covered material from the study of magnets and electricity, including:

- the nature of magnets and the origin of magnetic fields
- magnetic forces on currents and moving charges
- applications of magnetic forces—including electric motors.

9.1 Fundamentals of magnetism

Simple magnetism

If you put a paper clip or pin near a magnet, it will 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—a space in which magnetic effects can be noticed. 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.

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**.



Like magnetic poles repel each other, and 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 9.2).

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 may 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.

These observations soon give rise to some obvious questions.

- Can we obtain separate north and south magnetic poles, just as we can obtain separate positive and negative charges?
- Why is it that unmagnetised pieces of iron are attracted by either pole of a magnet?
- Can we change or destroy the magnetism of a magnet?
- How can we measure the strength of a magnet or, more particularly, that of a magnetic field?
- Are there other ways of producing magnetic effects?
- Are there materials other than iron and its alloys that show noticeable magnetism?

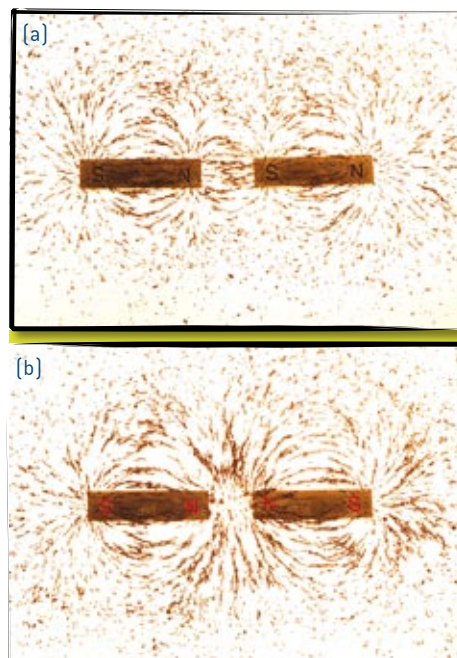


Figure 9.1 Iron filings sprinkled around (a) magnets with unlike poles together and (b) with like poles together. The patterns in the fields clearly show the attraction and repulsion.

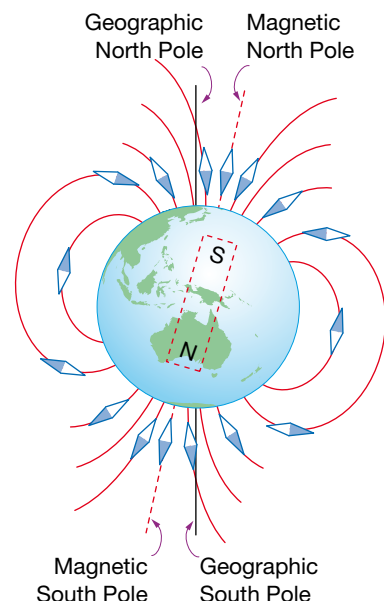


Figure 9.2 The Earth acts as though it has a south magnetic pole near the geographic north pole! The ‘magnetic north pole’ is the place to which the north end of a compass appears to point.

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.



PRACTICAL ACTIVITY 29

Magnetic field of a permanent magnet

- What causes the Earth's magnetic field? Why does it change slowly? What is its significance to us?

We will answer two of these questions here, and return to the others later in this chapter.

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**.



Figure 9.3 Magnets are always dipolar. When a magnet is broken, two new poles appear at the broken ends.

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.

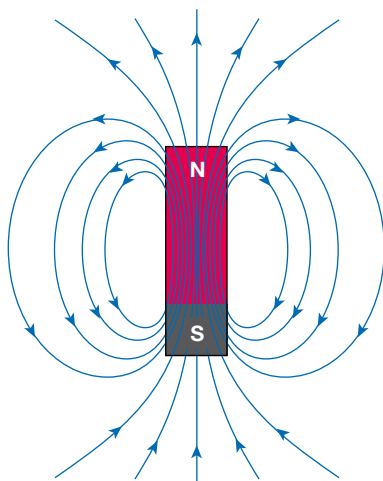


Figure 9.4 The field lines around and inside a bar magnet. The lines show the direction of the force on an (imaginary) single north pole.

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 measure the strength and direction of a magnetic field. We simply define the direction of the field as that of the direction of the magnetic force on the north pole of a magnet. The force on a south pole, then, is always in the direction opposite to the field. It is easy to see why a magnet will always try to align itself with the field. A **torque** (a turning force that will try to rotate an object) will be created if the two poles are not aligned with the field (see Figure 9.6). This will be true whether the magnet is a bar magnet hung on a string in the Earth's field, a tiny iron filing near a bar magnet, or a specially shaped and balanced magnet mounted in an oil-filled case—a magnetic compass.

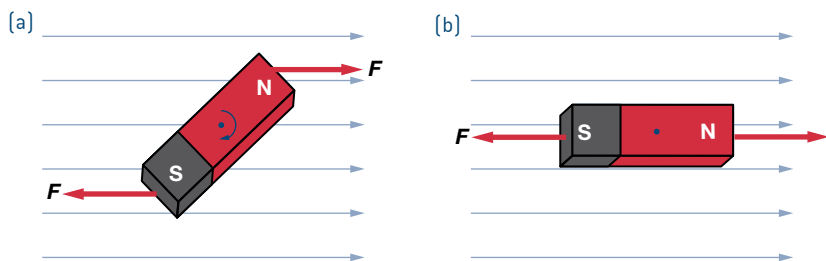


Figure 9.5 A strong magnet in the base induces each little leaf to become a magnet.

Figure 9.6 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 BC. 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 BC 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.)



Figure 9.7 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.

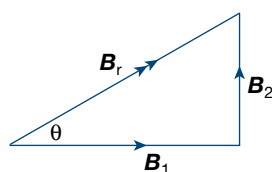


Figure 9.9 The two magnetic fields add as vectors, giving a resultant field to which the compass needle responds.

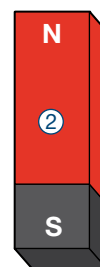


Figure 9.8 The compass needle points away from the north ends of the two magnets, but clearly magnet 1 is stronger than magnet 2 in this case.

★ Although we are not ready to define the strength of a magnetic field in absolute terms, we can certainly compare the relative strengths 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 9.8, 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.

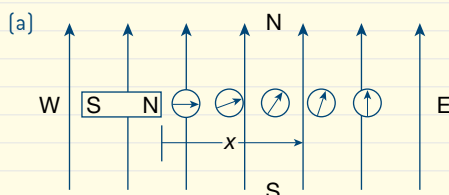
A simple piece of vector analysis enables us to compare the strengths of the two fields. The symbol \mathbf{B} is used to represent the vector magnetic field. In Figure 9.9, if \mathbf{B}_1 represents the field due to magnet 1 at the location of the compass and \mathbf{B}_2 represents that of magnet 2 at the same place, we can see that they can be added vectorially to give \mathbf{B}_r , the resultant field to which the compass will respond.

The direction θ of the resultant field is given by $\tan\theta = B_2/B_1$, so that $B_2 = B_1 \tan\theta$. In Figure 9.9, θ is 30° . As $\tan 30^\circ = 0.58$, we can see that the strength of B_2 is given by $B_2 = B_1 \times 0.58$. That is, the field B_2 due to magnet 2 is 58% of the field B_1 of magnet 1. (If the angle had been 45° , $\tan\theta = 1$ and so $B_2 = B_1$. The two fields would be equal.)

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!

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 9.8. The bar magnet is aligned east–west and the compass is moved along a line extending lengthwise from the magnet. At any position, the direction 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. $\theta = 45^\circ$), 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 9.10b 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.

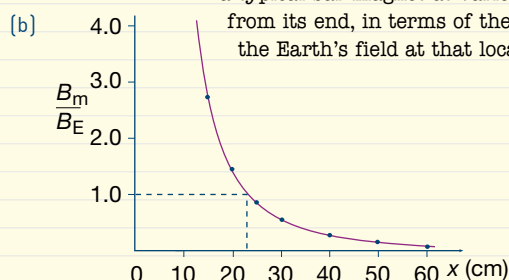


Figure 9.10 (a) The experimental set-up and (b) the graph of B_m/B_E , the ratio of the strength of the field of the magnet compared to the Earth's field, against distance in cm from the magnet. Note that in this case the field of the magnet is equal to the Earth's field at about 23 cm from the magnet.



9.1 summary

Fundamentals of magnetism

- 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.
- An unmagnetised piece of iron will have magnetism induced in it when placed in an external magnetic field.
- 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.

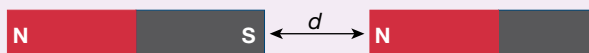
- When a magnet is placed in a magnetic field, magnetic forces cause a torque which tends to align it with the field, the north end of the magnet pointing in the direction of the field.
- The magnetic field, \mathbf{B} , is a vector quantity. Vector addition enables us to add fields and to compare the magnitudes of different fields.



9.1 questions

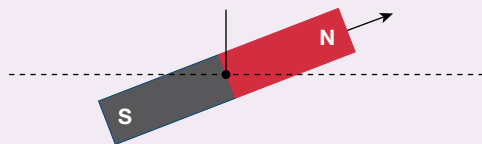
Fundamentals of magnetism

- 1 Explain how a magnetic field might be detected at any point in space.
- 2 The following diagram shows two bar magnets separated by a distance d . At this separation, the magnitude of the magnetic force between the poles is equal to F . Which of the following is true if d is increased?



- A An attractive force greater than F will exist between the poles.
- B A repulsive force greater than F will exist between the poles.
- C An attractive force less than F will exist between the poles.
- D A repulsive force less than F will exist between the poles.

- 3 The following diagram shows a bar magnet suspended at its midpoint by a light wire. Which of the following is true?



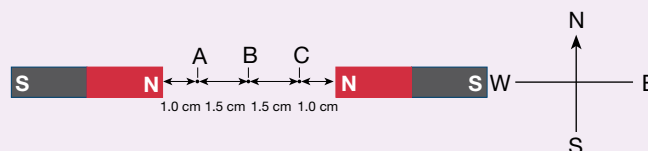
- A** The magnet is in the northern hemisphere and its N end is pointing towards the geographical south pole.
B The magnet is in the southern hemisphere and its N end is pointing towards the geographical north pole.
C The magnet is in the northern hemisphere and its N end is pointing towards the geographical north pole.
D The magnet is in the southern hemisphere and its N end is pointing towards the geographical south pole.
- 4 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** They are an inherent part of its atomic structure.
B They depend on the shape and dimensions of the material.
C They depend on the arrangement of the atoms in the crystal lattice.
- 5 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.
- 6 Explain why a magnet will always try to align itself with the direction of the existing magnetic field.
- 7 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 10^3B ?

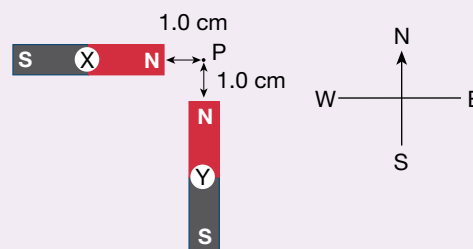
The following information applies to questions 8 and 9. Two strong bar magnets which produce magnetic fields of equal strength are arranged as shown.



- 8 Ignoring the magnetic field of the Earth, describe the approximate direction of the resulting magnetic field at:
- a** point A
b point C.
- 9 What would be the magnitude of the magnetic field at point B?

The following information applies to questions 10 and 11.

Two identical bar magnets X and Y are located as shown.



- 10 Ignoring the magnetic field of the Earth, describe the direction of the resultant magnetic field at point P for magnets with the following relative fields:
- a** magnet X = B , magnet Y = $10^{-3}B$
b magnet X = B , magnet Y = 10^3B
c magnet X = B , magnet Y = $\sqrt{3}B$
- 11 The magnetic field at point P due to magnet X is equal to $4B$, while the field at P due to magnet Y is equal to $3B$. What is the magnitude of the resultant magnetic field at point P?

9.2 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!



Figure 9.12 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.



Figure 9.11 Hans Christian Oersted discovered the magnetic effect of an electric current in 1820.

Physics file

By the early 17th century, ‘static’ electricity and ‘magnetostatics’ had become interesting curiosities, but did not attract much serious attention from scientists. However, 70 years before Oersted, Benjamin Franklin had aroused greater interest from scientists by showing that lightning was actually electricity on a grand scale. But in 1800 Galvani discovered that electricity could be produced by the chemical action of different metals, and just 1 year later Alessandro Volta invented the ‘voltaic pile’. This produced *current* electricity from piles of metal disks separated by acid-soaked paper—the forerunner of our modern batteries. Unlike the electrostatic machines, Volta’s piles could deliver quite a lot of power—enough to make a wire glow red hot or to produce a brilliant white light from a carbon arc, for example. Even more interestingly perhaps, Humphrey Davy (director of the Royal Institution in London), by passing an electric current through molten soda and potash, showed that they contained two new metals never seen before—sodium and potassium. It would have been these types of phenomena that Oersted was demonstrating in his lectures.

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:

$$B = \frac{KI}{r}$$

where B = the strength of the magnetic field

I = the current

r = the distance from the wire

K = a constant which depends on the medium around the current.

It is easy to remember which way the magnetic field around a current points 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. (It is also called a right-hand screw rule because if you turn a normal right-handed screw in the direction of your fingers, it moves in the direction of your thumb.)

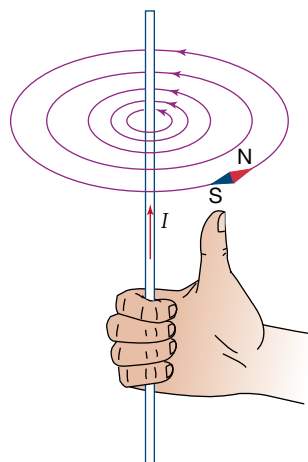


Figure 9.13 The direction of the field around a straight wire is given by the right-hand grip rule.

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 flows in the wire, the wire moves—there is a magnetic force acting on it. Further investigation reveals that the force is maximum if 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.

Another simple **right-hand force rule** helps remember these relative directions. Orient your right hand so that your thumb points in the direction of the current (as for the previous rule) and your fingers, held straight this time, are in the direction of the external field (**fingers = field**). The force is then in the direction we would normally push: out from our palm.

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 9.13) 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 force on a current in a magnetic field (Figure 9.15). *This force is the result of putting a current in a magnetic field*. Don't 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 9.16).



PRACTICAL ACTIVITY 30

Direction of induced current in a wire

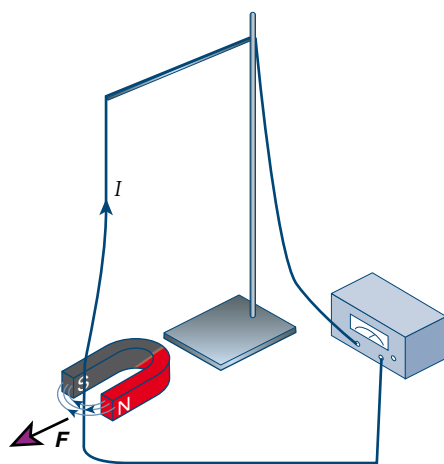


Figure 9.14 When current flows in the flexible wire between the poles of a strong U magnet, the wire experiences a force at right angles to the current and to the field.

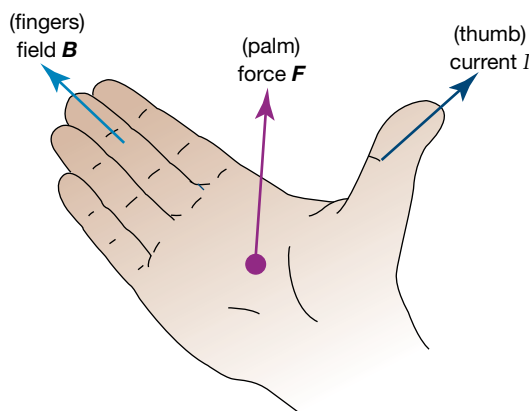


Figure 9.15 The right-hand force (or palm) rule gives the direction of the force, F , on a current, I , in a magnetic field, B . The thumb represents the direction of the current, the fingers represent the direction of the field, and the direction of the force is out from the palm of the hand, as you would push.

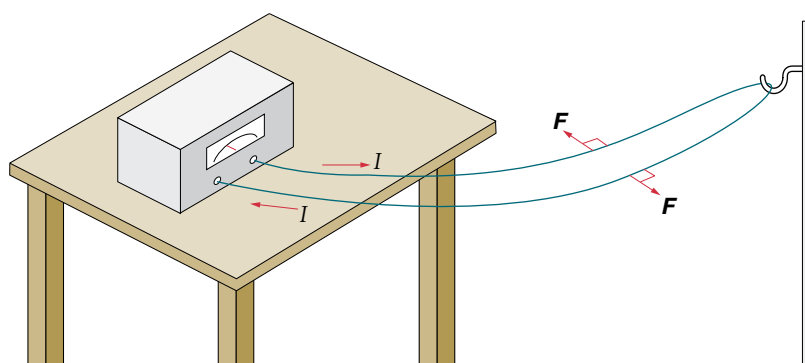


Figure 9.16 Two wires with a current flowing in opposite directions experience a repulsive force. You can repeat this experiment with a length of flexible wire connected to a power supply and looped over a hook on the wall. Have the wires about 2 cm apart and at least 2 metres long, and use the maximum 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.



9.2 summary

The foundations of electromagnetism

- 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.
- Anelectriccurrentplacedinaperpendicularmagnetic field experiences a magnetic force perpendicular to both the current and the field. The direction of the force is given by the right-hand force or palm rule.
- It is important not to confuse the two right-hand rules. The grip rule gives the direction of the field produced by a current, the palm rule gives the direction of the force on a current-carrying wire placed in an external field.
- As a current produces a magnetic field and also experiences a force when placed in a magnetic field, there is a magnetic force between two parallel currents.

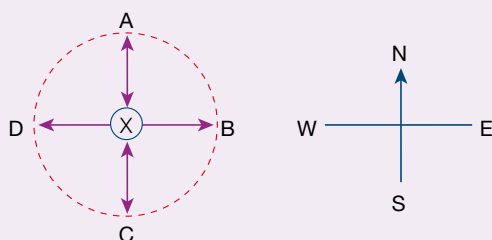


9.2 questions

The foundations of electromagnetism

The following information applies to questions 1–4.

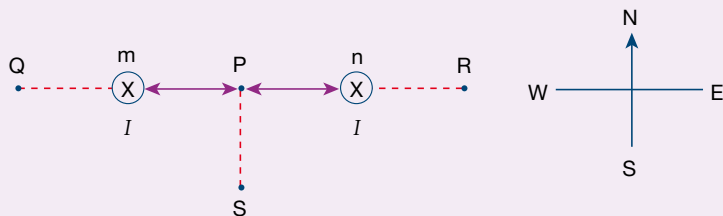
The following diagram shows a cross-sectional view of a long, straight, current-carrying conductor, with its axis perpendicular to the plane of the page. The conductor carries an electric current into the page.



- 1 What is the direction of the magnetic field produced by this conductor at each of the points A, B, C and D?
- 2 The direction of the current in the conductor is now changed so that it is carried out of the page. What is the direction of the magnetic field produced by this conductor at the four points A, B, C and D?
- 3 For the current described in Question 1, what is the direction of the magnetic force exerted on it by the Earth's magnetic field?
- 4 For the current described in Question 2, what is the direction of the magnetic force exerted on it by the Earth's magnetic field?

The following information applies to questions 5–8.

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.



- 5
 - 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?
- 6 At which of the points Q, R and S, could the resultant magnetic field due to the currents and the Earth be zero?
- 7 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?
- 8
 - a With conductor n carrying current into the page, what is the direction of the magnetic force on conductor n due to the field produced by conductor m?
 - b With conductor n carrying current out of the page, what is the direction of the magnetic force on conductor n due to the field produced by conductor m?

9.3 Currents, forces and fields

The force on a wire carrying a current in a magnetic field (**B**) 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. $F \propto Il$.

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 **B**? In discussing the gravitational field **g** and the electric field **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 intensity of the magnetic field is then the force per unit length of unit current. That is:

$$B = \frac{F}{Il}$$

We can write this relationship as $F = IlB$, and so we see that the constant in our earlier proportionality is *B*, the strength of the magnetic field.



The force on a wire carrying a current in a magnetic field is proportional to the current, the length of wire in the field, and the strength of the field:

$$F = IlB$$

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 (in newtons) on a current (in amperes) in a field of length *l* (in metres). We say that the unit for field strength will be the field that produces a force of 1 newton on a 1 metre length of a current of 1 ampere. The unit for magnetic field then becomes the newton per amp metre ($\text{N A}^{-1} \text{m}^{-1}$). 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 ($1 \text{ T} = 1 \text{ N A}^{-1} \text{m}^{-1}$)

The Earth's magnetic field at the surface is around $5 \times 10^{-5} \text{ T}$. 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.)

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.

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.

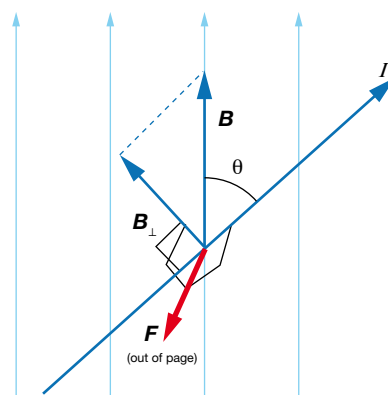


Figure 9.17 The magnitude of the force on a current depends on the component of the field at right angles to the current. Another way of writing the expression is $F = IlB_{\perp}$ where B_{\perp} is the component of the field at right angles to the current. In fact, of course, $B_{\perp} = B \sin \theta$.

Physics file

The three quantities force, length and field are all vectors (\mathbf{F} , \mathbf{l} and \mathbf{B}). To express the relationship between them fully we need what is called 'vector cross multiplication'. If we write $\mathbf{F} = I \times \mathbf{B}$, this is taken to mean that the magnitude of vector \mathbf{F} is the product $IlB\sin\theta$, where θ is the angle between \mathbf{l} and \mathbf{B} , and the direction is at right angles to both \mathbf{l} and \mathbf{B} and in the sense given by the right-hand rule.



PRACTICAL ACTIVITY 31

Force on a current-carrying conductor

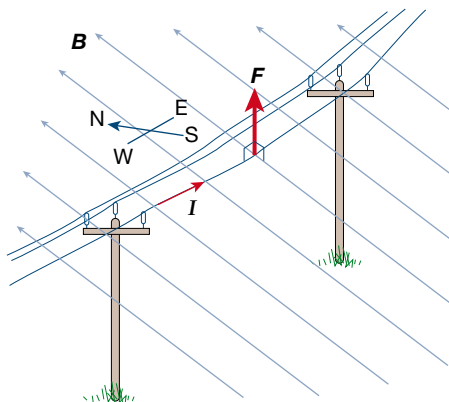
Table 9.1 Some typical magnetic fields

The surface of a neutron star	10^8 T
Largest artificially produced pulsed fields	10^3 T
Very strong electromagnets and 'supermagnets'	1–20 T
Magnetic storms on the Sun	1 T
Alnico and ferrite magnets	10^{-2} –1 T
The Earth's surface	5×10^{-5} T
Interstellar space	10^{-10} T
Smallest value achieved in a magnetically shielded room	10^{-14} T

There is often some confusion around the name and units of \mathbf{B} . While we have talked of 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. The name given to \mathbf{B} is more correctly *magnetic induction* or *magnetic flux density*, but for simplicity we will continue (in line with common practice) to just refer to the *strength of the magnetic field*.

Worked example 9.3A

Determine the magnetic force, due to the Earth's magnetic field, on a suspended power line running east–west and carrying a current of 100 A, as shown in the diagram.



Solution

As the line is running east–west, the current and field are at right angles. If we assume the Earth's magnetic field is 5×10^{-5} T, the magnitude of the force on 1 m of this power line is given by:

$$\begin{aligned}
 F &= IlB \\
 &= 100 \times 1 \times 5 \times 10^{-5} \\
 &= 5 \times 10^{-5} \text{ N}
 \end{aligned}$$

The direction of the force will be up and a little to the south. In practice, the current in a power line is normally AC and so the force would alternate in direction.

We have seen that an electric current that is placed in a magnetic field at right angles to the current experiences a force. 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 \mathbf{F} , \mathbf{l} and \mathbf{B} is:



$$F = ILB\sin\theta$$

where θ is the angle between I and B .

Worked example 9.3B

If the 100 A power line in Worked example 9.3A runs north–south instead of east–west, and the dip angle of the Earth’s field is 40° , what would be the force on the line? [The dip angle is the angle of the field below the horizontal.]

Solution

This time the field and current are not at right angles and so we need to include $\sin\theta$. The dip angle is the angle of the field below the horizontal and so it is also the angle between the current and the field.

$$\begin{aligned} F &= ILB\sin\theta \\ &= 100 \times 1 \times 5 \times 10^{-5} \times \sin 40^\circ \\ &= 3.2 \times 10^{-3} \text{ N} \end{aligned}$$

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.



9.3 summary

Currents, forces and fields

- The magnitude of the force on a wire carrying a current I in a magnetic field B is given by:

$$F = ILB\sin\theta$$
 where I is the length of current in the field and θ is the angle between the current and field.
- The unit for magnetic field, 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.



9.3 questions

Currents, forces and fields

The following information applies to questions 1–3.

An east–west power line of length 100 m is suspended between two towers. Assume that the strength of the magnetic field of the Earth in this region = 5.0×10^{-5} T.

- Calculate the magnetic force on this power line if it carries a current of:
 - 80 A from west to east
 - 50 A from east to west.
- Assuming that $g = 10 \text{ N kg}^{-1}$ and that the total mass of the power line is 50 kg, calculate the value of the

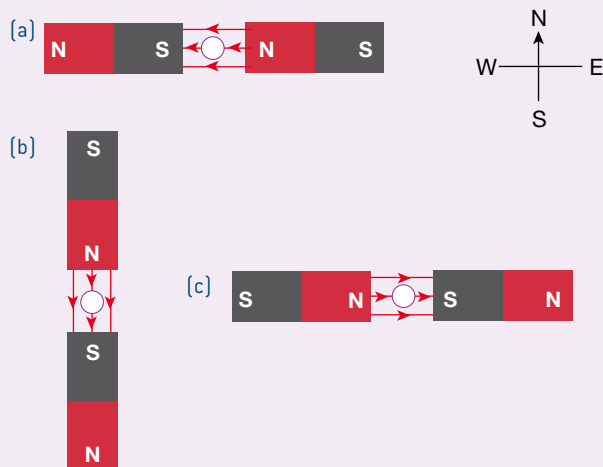
ratio of the weight to the magnetic force on the power line when carrying a current of 100 A.

- 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:
 - greater than before?
 - the same as before?
 - 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, \mathbf{B} , of these magnets, and the currents, \mathbf{I} , are mutually perpendicular in all cases.



- 4 For diagram (a), calculate the magnitude and direction of the magnetic force on a 5.0 cm section of conductor with:

- a $I = 2.0$ A into the page, $B = 2.0 \times 10^{-3}$ T
b $I = 1.0$ A out of the page, $B = 2.0 \times 10^{-3}$ T

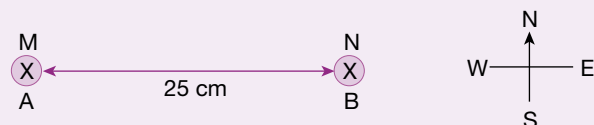
- 5 For diagram (b), calculate the magnitude and direction of the magnetic force on a 1.0 mm section of conductor for:

- a $I = 3.0$ A into the page, $B = 0.50$ T
b $I = 3.0$ A out of the page, $B = 1.0$ T

- 6 For diagram (c), calculate the magnitude and direction of the current that would result in a magnetic force on a 2.0 mm section of conductor of:

- a 8.0×10^{-3} N south, for $B = 0.10$ T
b 2.0×10^{-2} N north for $B = 0.50$ T

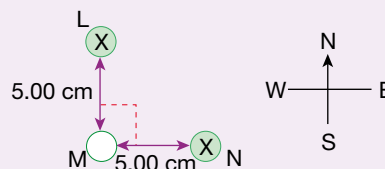
- 7 The following diagram shows two parallel, current-carrying conductors, M and N, located at points A and B, each carrying currents of 5.0 A into the page.



The magnetic field at point B due to M is equal to 4.0 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 1.0×10^{-3} T north-west.

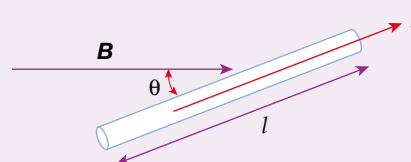


- 8 Calculate the magnitude and direction of the magnetic force per metre on M for:

- a a current of 2.0 A flowing through M into the page
b a current of 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 \mathbf{B} and \mathbf{I} is θ . Calculate the magnitude of the magnetic force on this conductor for:

- a $B = 1.0$ mT $I = 1.0$ mA $l = 2.0$ mm $\theta = 90^\circ$
b $B = 1.0$ T $I = 10$ A $l = 5.0$ cm $\theta = 0^\circ$
c $B = 0.10$ T $I = 5.0$ A $l = 1.0$ mm $\theta = 30^\circ$

9.4 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.

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 representing the tail feathers of an arrow going away from you, and the dots the point of the arrow coming towards you.) So the field around the loop could be drawn as in Figure 9.19, 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, their fields all 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.

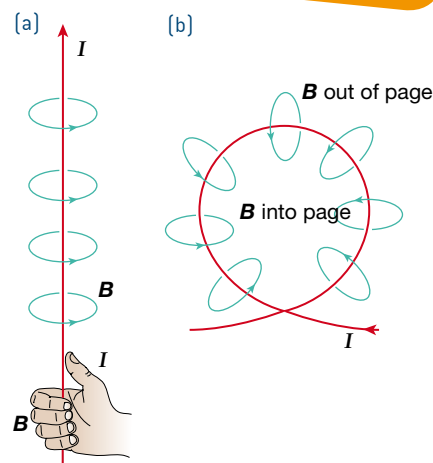


Figure 9.18 (a) The field around a straight current-carrying wire is circular. (b) The wire is bent into a clockwise loop. The field is now into the page in the centre and out of the page around the outside.

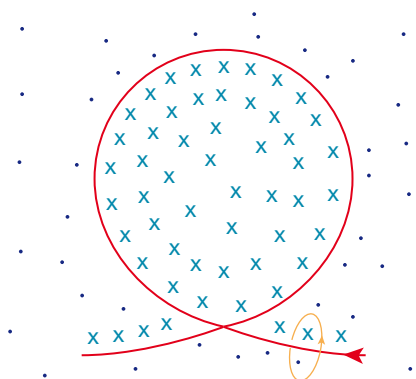


Figure 9.19 Field lines coming out of the page are shown by dots, those going into the page by crosses. The density of the crosses inside the loop will be greater than the density of the dots outside, suggesting a stronger field inside the loop.

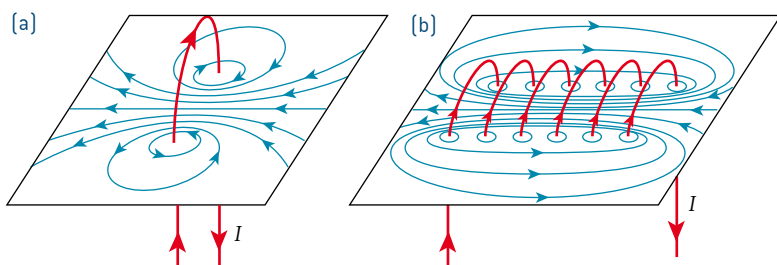


Figure 9.20 (a) The field lines around a single current loop. (b) The field lines become much more concentrated inside a solenoid.

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



PRACTICAL ACTIVITY 32

Magnetic field of a solenoid

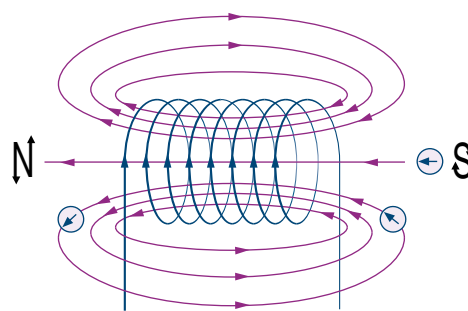


Figure 9.21 This solenoid has an effective 'north' end at the left and a 'south' end at the right. The compass points in the direction of the field lines, towards the south pole of another magnet. A simple way to remember which end of a solenoid is which pole is to put arrows on the letters N and S as shown. The arrows indicate the direction of the current as seen from that end.

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? Could it be that, in some way, magnets had tiny electric currents in them?

Physics in action

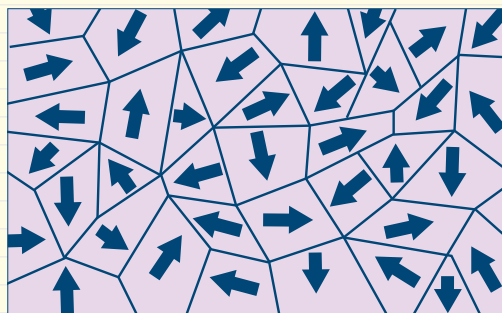
Magnetic domains

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.

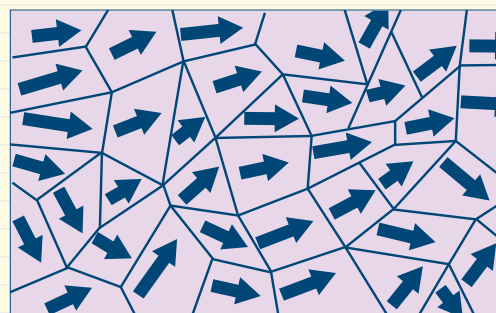
Each domain has a very strong (but tiny) magnetic field, but if all the domains are oriented differently, the overall effects will cancel out. When the iron is put into a magnetic

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.



non-magnetic



magnetic

Figure 9.22 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.

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* (see Physics in action, page 342).



Figure 9.23 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.

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 **a**luminium, **n**ickel and **c**obalt, 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.

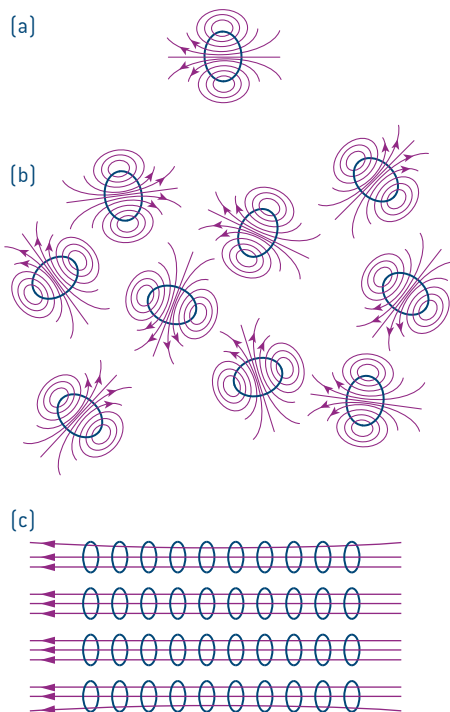


Figure 9.24 (a) A little atomic electromagnet with its magnetic field. (b) Many little electromagnets oriented randomly produce no external field. (c) Many little electromagnets aligned with their fields parallel produce a large external field.



PRACTICAL ACTIVITY 33

Investigating electromagnetic induction

The Earth: One huge electromagnet

It is common to see pictures illustrating the Earth's magnetic field with a large bar magnet, with its 'S' pole uppermost, inside the Earth. Of course, there is no huge bar magnet inside the Earth. The magnetic field arises from huge electric currents deep inside the Earth. But where do the electric currents come from and how do we know anyway?

Perhaps the clearest indication that the magnet inside the Earth is not a 'permanent' one is that the field of the Earth varies with time. Most map users know that even over a few years the *magnetic declination* (the angle between true north and magnetic north) changes slightly. This is because the Earth's field is actually in a constant state of turmoil! Over thousands of years it changes very considerably. In fact, in the last 1000 years its average strength has dropped by about 40%. At the present rate of change, it would disappear within 2000 years. However, this sort of change is quite normal for the Earth's magnetic field.

When molten rock sets hard, the iron in it becomes permanently magnetised as a result of being in the Earth's field. By looking at the magnetism set in rocks formed throughout the Earth's geological history, geophysicists are able to deduce that the Earth's magnetic field has undergone complete reversals of polarity at rather irregular intervals. In the last 3.6 million years, there have been about 15 changes of polarity, the last being about 730 000 years ago. It is thought that the present rate of decrease in the strength of the field may indicate that we could be due for another reversal within the next few thousand years.

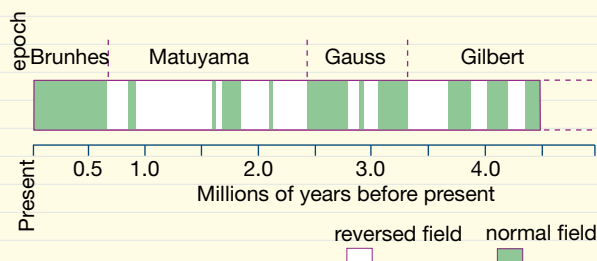


Figure 9.26 Over the last 5 million years the direction of the Earth's magnetic field has reversed at least 20 times.

If we were to have a reversal, would we lose the magnetic shield that protects us from the high-energy particles from the Sun and cosmic rays? No one really knows. However, the good news is that the fossil record shows that life on Earth has survived reversals in the past, although possibly with some damage.

Not only does the Earth's field change in time, but it also varies irregularly over the Earth's surface. At present the Earth's field varies from around 25 μT in South America to 65 μT near the magnetic poles. In Australia, it varies from around 45 μT in the north to 60 μT in the south. These irregularities are taken as an indication that the process that generates the field is itself a rather turbulent one. Geophysicists now believe they have a fairly good picture of this process. They see the Earth as a self-sustaining dynamo-electromagnet. In other words, it produces the magnetic field from huge electric currents generated in the swirling molten iron that makes up the Earth's core. These currents are themselves generated by the magnetic field they create.

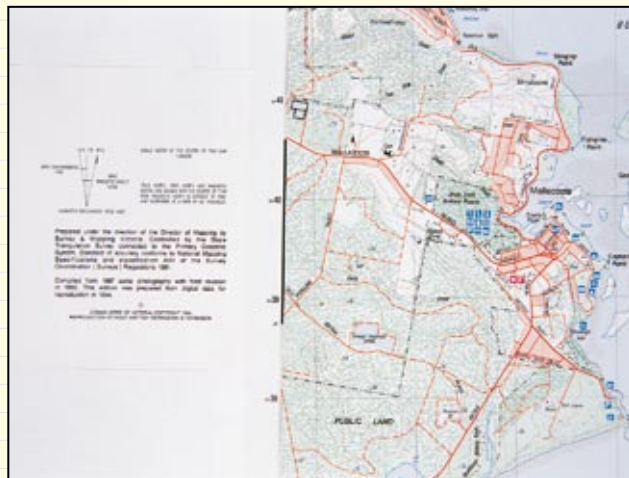


Figure 9.25 Magnetic declination changes gradually. The present declination and the annual change are recorded on official maps.



9.4 summary

Magnetic fields around currents, magnets and atoms

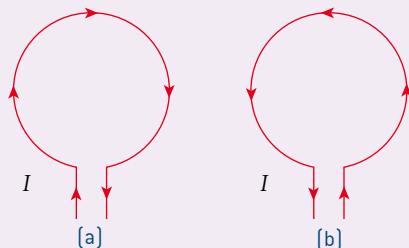
- The magnetic field produced by a single loop of current is perpendicular to the loop, and is stronger inside the loop than outside.
- The magnetic field of a solenoid is much stronger inside than outside.
- The field around a loop or solenoid is a dipole field, like that of a bar magnet.
- A soft iron 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).



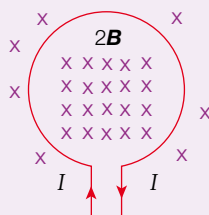
9.4 questions

Magnetic fields around currents, magnets and atoms

The following information applies to questions 1 and 2. The diagrams below show a current loop with its plane parallel to the page.



- 1 What is the direction of the magnetic field inside the loop in diagram (a)?
- 2 What is the direction of the magnetic field inside the loop of diagram (b)?
- 3 Explain why the magnetic field strength inside a current loop is stronger than an equivalent point outside a loop.
- 4 The following diagram shows a loop carrying a current I which produces a field \mathbf{B} in the centre of the loop. It is in a region where there is already a steady field of \mathbf{B} (the same as that due to I) directed into the page, so that the resultant field is $2\mathbf{B}$.

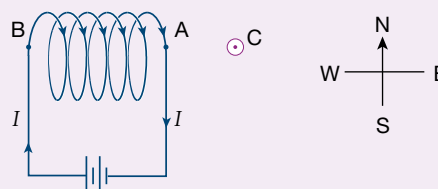


What would the magnitude and direction of the resultant field at the centre of the loop become in each of the following cases?

- 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 5–8.

The following 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.



- 5 If the direction of the magnetic force on this conductor is north, which point (A or B) represents the north pole of this solenoid?
- 6 If the direction of the current through the solenoid is reversed, what is the subsequent direction of the magnetic force on C?
- 7 The direction of the current through the solenoid is returned to its original direction and the direction of the current through C is reversed (it is now into the page). What is the direction of the magnetic force on C?
- 8 Which one or more of the following would cause this solenoid to produce a magnetic field of greater strength than before?
 - A Increase the value of I .
 - B Insert a copper cylinder inside the solenoid.
 - C Insert a soft iron cylinder inside the solenoid.
- 9 Which one or more of the following produces a dipole field?
 - A A long, straight, current-carrying conductor
 - B A current-carrying solenoid
 - C A permanent magnet
- 10 A solenoid cannot induce magnetism in a piece of:
 - A nickel
 - B cobalt
 - C aluminium
 - D gadolinium.

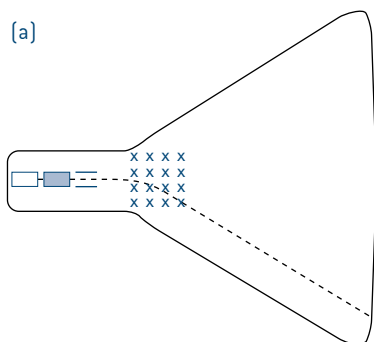
9.5

Forces on moving charges ☆

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 9.27b).

(a)



(b)

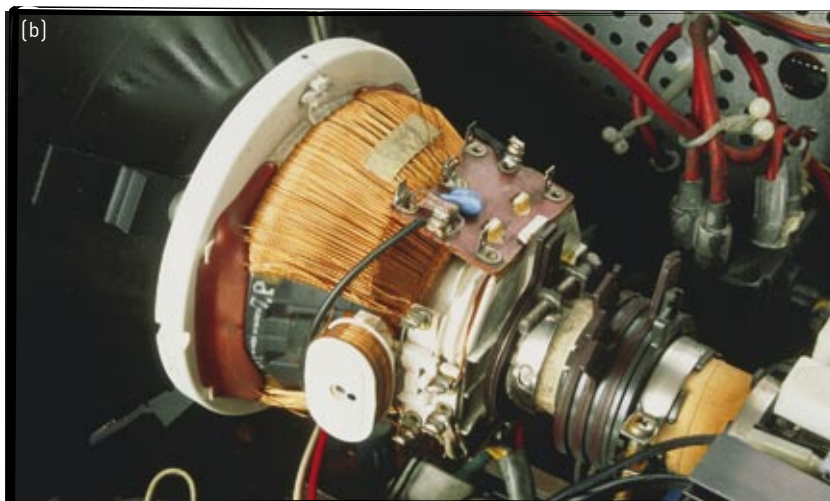


Figure 9.27 The deflection coils, or yoke, at the neck of a computer monitor cathode ray tube cause the beam to be deflected through a large angle, so the screen can be made reasonably short. The current in the coils varies rapidly in order to create the magnetic field which causes the beam to sweep 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.

Physics file

The force on a 1 m length of wire carrying a current I in a perpendicular field B will be given by $F = IB$ (taking $l = 1$). If in that 1 m of wire, the current is carried by n charges which each carry q coulombs of charge. The force F_q on each individual charge is then given by:

$$F_q = \frac{F}{n} = \frac{IB}{n}$$

If the charges are all moving at speed v , they will take $t = 1/v$ seconds to move through this 1 m length. The current will be given by $I = Q/t$, where Q is the total charge (nq) that moves through the 1 m length in time t . As this time $t = 1/v$, the current is given simply by $I = Qv$. So the expression for the force on a single charge q becomes:

$$F_q = \frac{IB}{n} = \frac{QvB}{n} = \frac{nqvB}{n} = qvB$$

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 force F_q on a charge q moving with velocity v perpendicular to a field B is given by:

$$F_q = qvB$$

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 expression becomes:

$$F_q = qvB \sin\theta$$

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!

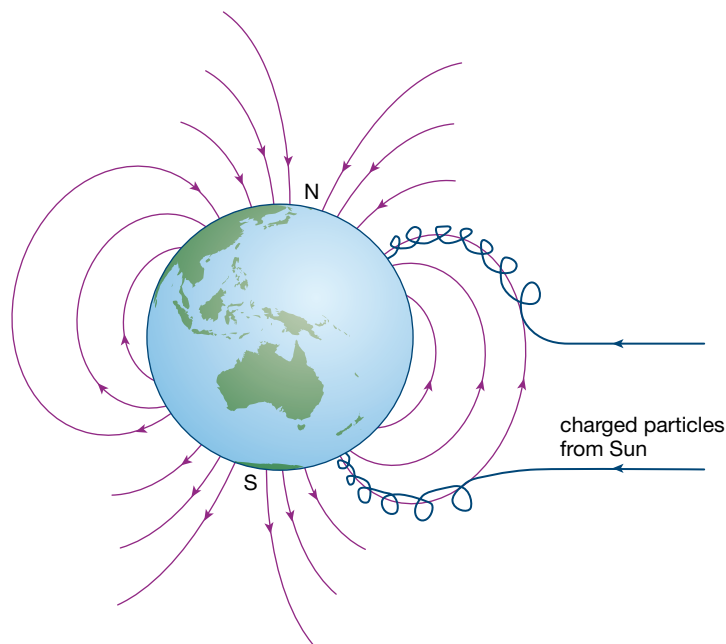


Figure 9.28 Charged particles from the Sun or deep space are trapped by the Earth's magnetic field, which causes them to spiral towards the poles. As they do this, they lose their energy and become harmless to the life on Earth.

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, B_{\perp} , the thumb this time represents the direction of the velocity, 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.

For more information see Chapter 13 'Synchrotrons and applications'.

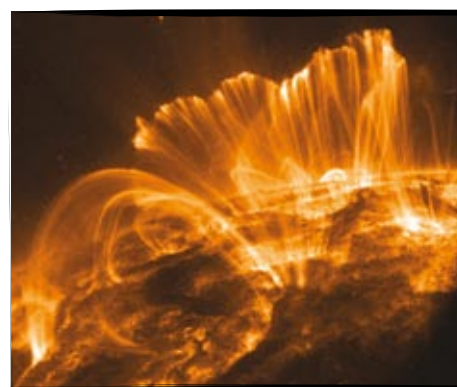


Figure 9.29 Solar flares are caused by huge ejections of gas from the Sun's surface. Because the gas is so hot it is ionised and so traces out a gigantic arc as the charged particles move in the Sun's enormous magnetic field. This flare is over 20 times the size of the Earth.

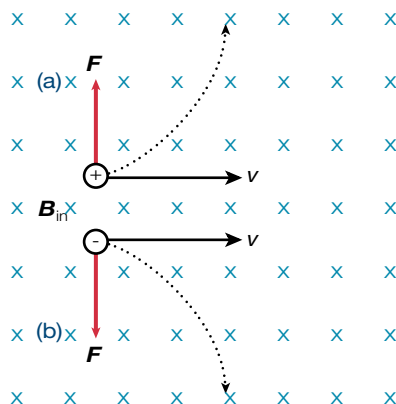
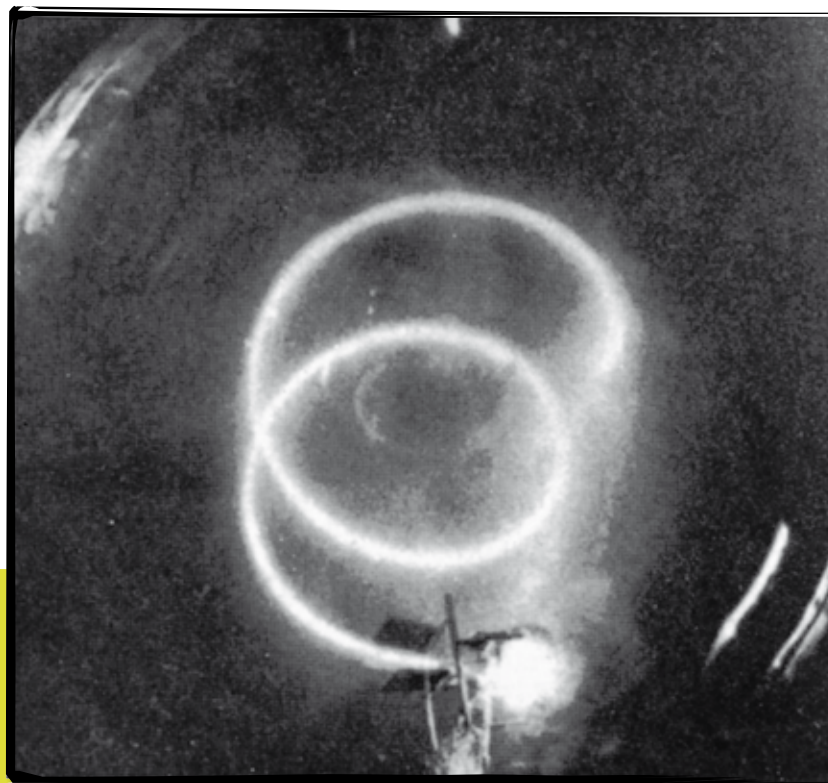


Figure 9.30 The direction of the force on a positive and a negative moving charge as they travel in a magnetic field directed into the page. Use the right-hand palm rule to confirm these directions for yourself.

For more information see Chapter 6 'Einstein's special relativity'.

Figure 9.31 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 exist. 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 name implies, 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 9.32 A simple mass spectrometer.

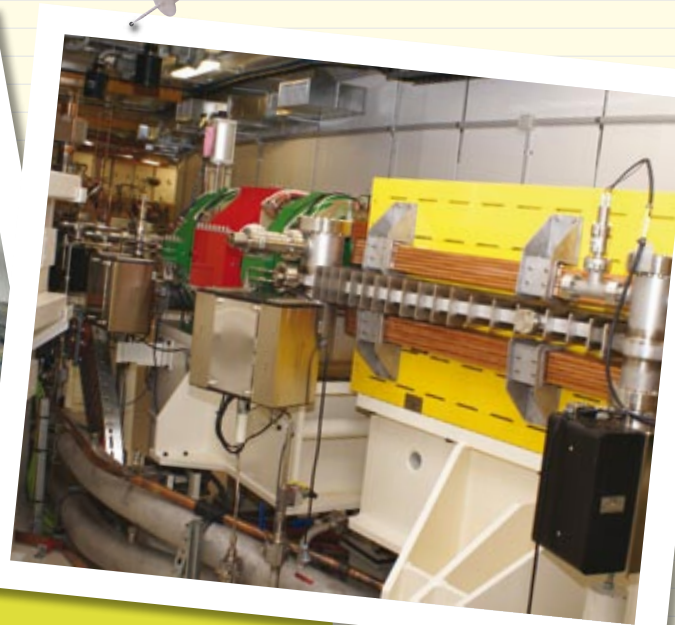


Figure 9.33 The Australian Synchrotron.



9.5 summary

Forces on moving charges

- Fundamentally, the magnetic force on an electric current is due to the force on the moving charges.
- The magnitude of the force \mathbf{F}_q on a charge q moving at speed v in a field \mathbf{B} is given by $F_q = qvB\sin\theta$. 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.



9.5 questions

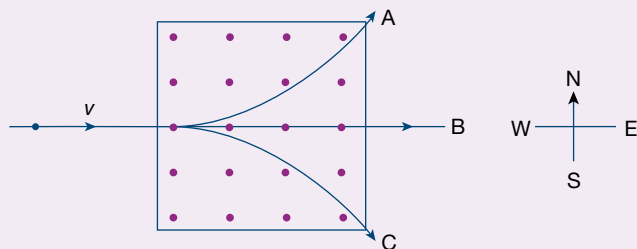
Forces on moving charges

- Which of the following quantities does not affect the magnitude of the magnetic force experienced by a particle moving through a magnetic field?
 - charge
 - velocity
 - mass
 - field strength
- 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:
 - perpendicular to the direction of the field
 - parallel to the direction of the field
 - at an angle θ to the field that is between 0 and 90° .

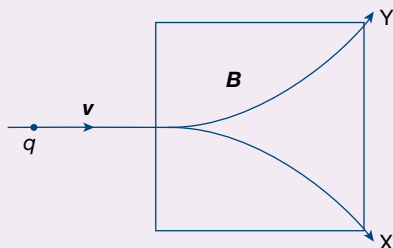


The following information applies to questions 3–6.

The following diagram shows a particle, with initial velocity \mathbf{v} , about to enter a uniform magnetic field, \mathbf{B} , directed out of the page.



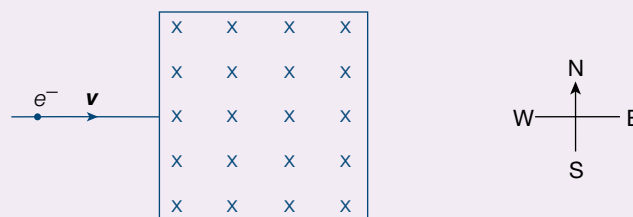
- 3 If this particle is positive:
 - a what is the direction of the force on this particle just as it enters the field?
 - b which of the paths A, B or C will this particle follow?
- 4 Which of the following statements is correct?
 - A The kinetic energy of this particle remains constant in the field.
 - 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?
- 7 The following diagram shows a particle with charge q entering a uniform magnetic field of strength \mathbf{B} . The initial velocity, \mathbf{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, \mathbf{B} is into the page.
- B q is negative, \mathbf{B} is out of the page.
- C q is positive, \mathbf{B} is out of the page.
- D q is negative, \mathbf{B} is into the page.

- 8 Which of the alternatives in Question 7 would result in the particle following path Y?

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, \mathbf{B} , directed into the page. The initial velocity, \mathbf{v} , of the electron is perpendicular to the field. The magnitude of the magnetic force on this electron in the field \mathbf{B} is F .



- 9 a What is the direction of the magnetic force \mathbf{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 $+2e$, with mass about 7500 times that of the electron) enters the field with a velocity \mathbf{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?

9.6 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.

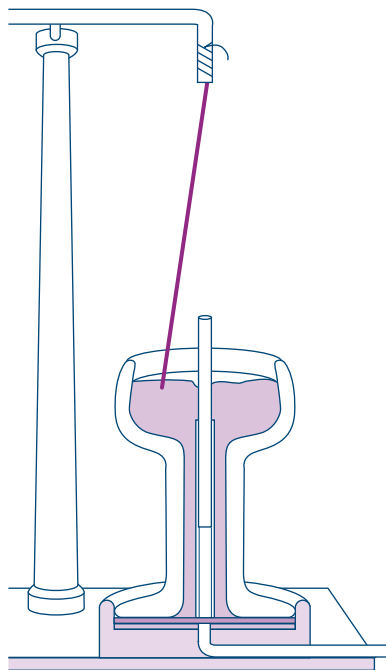


Figure 9.34 In 1821, Faraday built what could be called the first electric motor. A magnet was mounted vertically in a pool of mercury. A wire carrying a current hung from a support above. [The mercury provided a path for the current.] The magnetic field of the magnet spreads outwards from the top of 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.

The operating principle of all electric motors is simple. An electric current in a magnetic field experiences a force: $F = I\ell B$. 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, I , initially horizontal in a magnetic field, B (Figure 9.35). Current is flowing clockwise around this coil in the direction $A \rightarrow B \rightarrow C \rightarrow D$. 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 9.35a). 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 9.35b.



INTERACTIVE TUTORIAL
DC motors

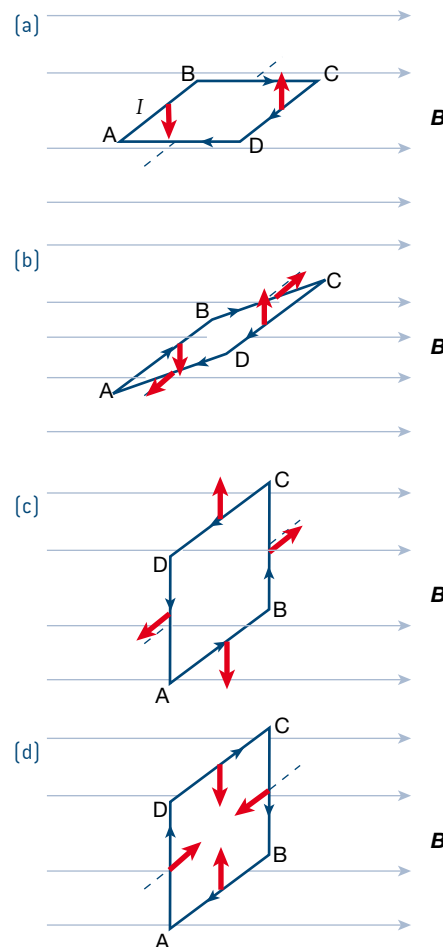


Figure 9.35 Forces acting on a current-carrying wire loop in a magnetic field.



PRACTICAL ACTIVITY 34
Current from an electric motor

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 force and the point of application. This is shown in Figure 9.36.

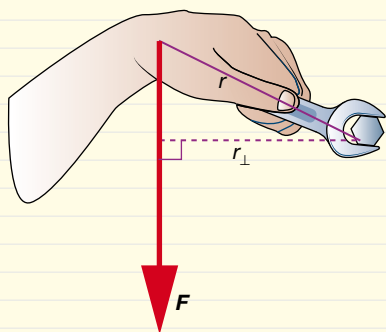


Figure 9.36

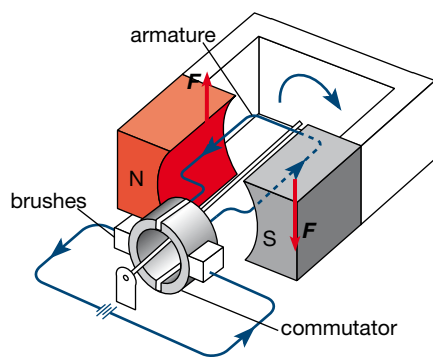


Figure 9.37 The commutator switches the direction of the current every half-turn in order to produce a torque which continues to rotate the coil in the same direction.

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 9.35c), 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 is reversed at this point. All the forces are then reversed, and the situation is as shown in Figure 9.35d. 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 9.37. 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.

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**.

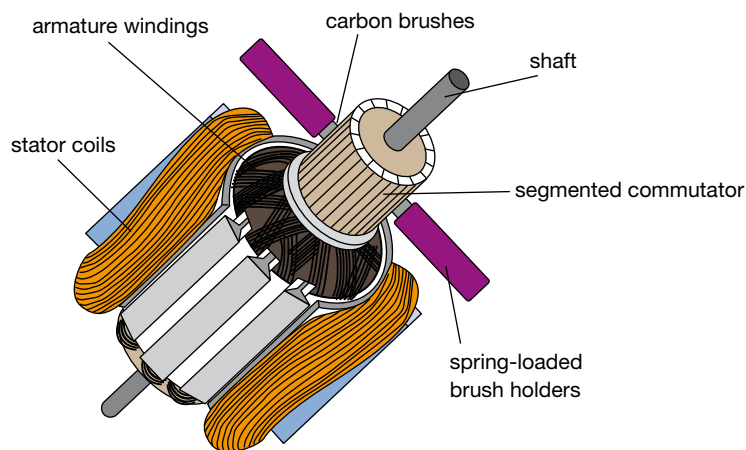


Figure 9.38 A typical universal electric motor, showing the main components. Some motors would have additional stator coils. The commutator feeds current to the armature coils in the position where most torque will be experienced. This type of motor will operate on either DC or AC power.

Worked example 9.6A

The coil shown in Figure 9.35a is 4.0 cm square and consists of 20 turns of wire and carries a current of 5.0 A in a field of 0.80 T.

- What are the forces on each side?
- What is the torque on the coil?

Solution

- 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 given by $F = nIlB$, where $n = 20$ turns, $I = 5.0$ A, $l = 0.040$ m and $B = 0.80$ T.

So:

$$F = 20 \times 5.0 \times 0.040 \times 0.80 \\ = 3.2 \text{ N}$$

The force on AB is down and on CD it is up.

- The torque on one arm of the coil is given by $\tau = rF$, where r is 2.0 cm (0.020 m) and F is 3.2 N.

The total torque will be twice this as both arms contribute the same torque

So:

$$\tau = 2 \times 0.020 \times 3.2 \\ = 0.13 \text{ N m}$$

Generally speaking, the higher the torque obtainable in an electric motor the better. 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.

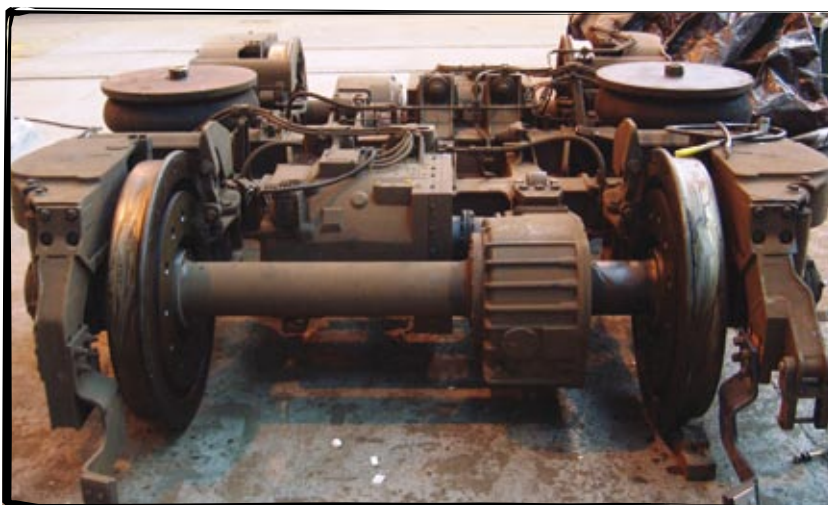


Figure 9.40 This is the power bogie of a modern Melbourne train. The three phase AC motor [centre left] drives the wheels via the gearbox (centre right). A complex electronic system converts the 1500 V DC from the overhead lines to variable frequency, variable voltage AC current as required by the speed 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

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.

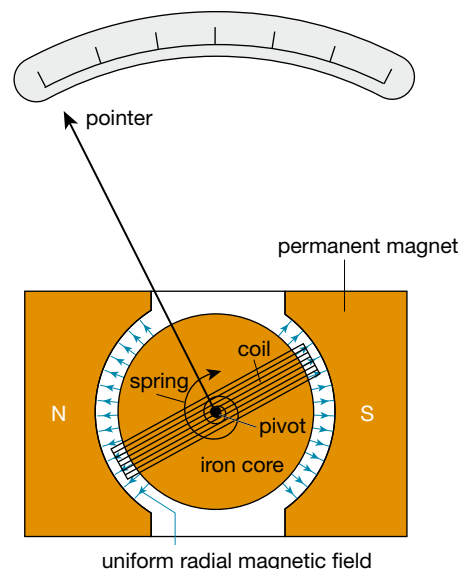


Figure 9.39 The movement of a galvanometer. The shaped pole pieces and the iron core in the centre intensify the field as well as making it more uniform. This is important if the meter is to be accurate.

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.



9.6 summary

Electric motors

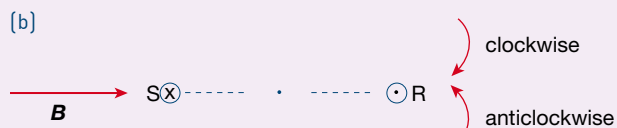
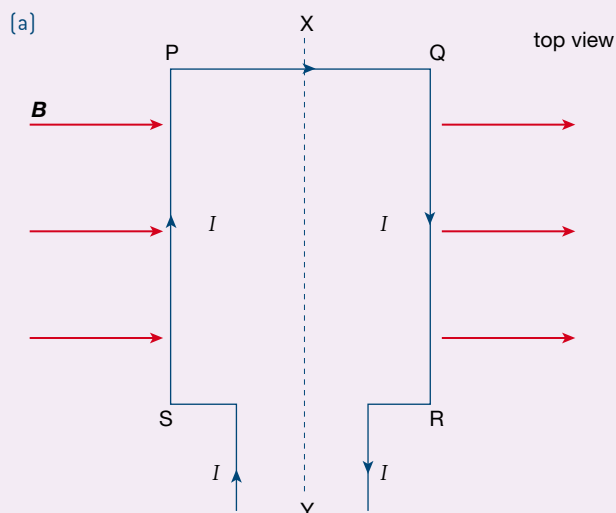
- An electric motor relies on the magnetic force on a current-carrying loop in a magnetic field:
 $F = I l B$
- There is a torque on the loop whenever its plane is not perpendicular to the field.
- The loop keeps rotating because the direction of current, and hence torque, is reversed each half turn 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.



9.6 questions

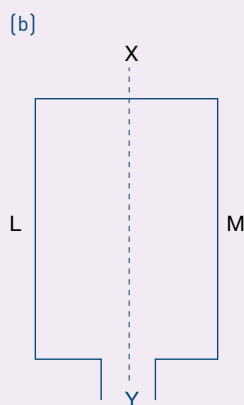
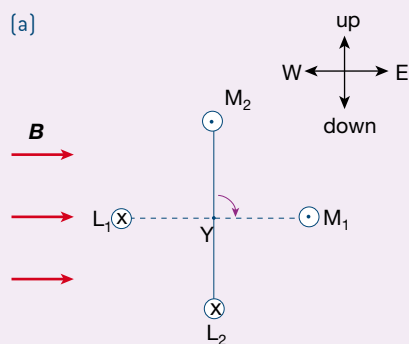
Electric motors

The following information applies to questions 1–5. Diagram (a) below depicts a top view of a current-carrying loop in an external magnetic field \mathbf{B} . Diagram (b) is the corresponding cross-sectional view as seen from Y. The following data applies to this diagram: $B = 0.10 \text{ T}$, $PQ = 2.0 \text{ cm}$, $PS = QR = 5.0 \text{ cm}$, $I = 2.0 \text{ A}$.



- Calculate the magnitude and direction of the magnetic force acting on side PS.
- Calculate the magnitude and direction of the magnetic force acting on side QR.
- What is the magnitude of the force on side PQ?
- This loop is free to rotate about an axis through XY. In what direction would this loop rotate? (As seen from Y.)
- 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. Diagram (a) below 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):

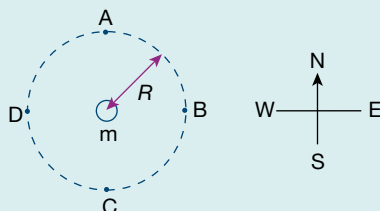


- 6 When LM is aligned horizontally ($L_1 - M_1$), what is the direction of the magnetic force on:
 - a side L?
 - b side M?
- 7 In what direction will the loop rotate?
- 8 When LM is aligned vertically ($L_2 - M_2$):
 - a what is the direction of the magnetic force on side L?
 - b what is the direction of the magnetic force on side M?
 - c what is the 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 review

The following information applies to questions 1–7.

The diagram below shows a conductor m with its axis perpendicular to the plane of the page. The local magnetic field strength of the Earth is $5.0 \times 10^{-5} \text{ T}$. A current, I , flowing through the conductor produces a magnetic field of magnitude $50 \mu\text{T}$ at a distance R .



- 1 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

- 2 What is the magnitude and direction of the magnetic field produced by this conductor at point B, if the resultant field at this point is zero?

- 3 What is the direction of the electric current responsible for producing this field?

- 4 What is the magnitude and direction of the resultant magnetic field at point B for a current, I , flowing through this conductor out of the page?

- 5 A current, I , flows through this conductor out of the page. What is the magnitude of the resultant field at point A?

- 6 What is the direction of the resultant magnetic field at point C for a current, I , flowing into the page?

- 7 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?

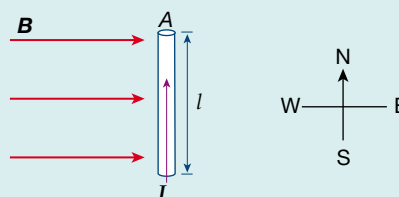
- A $5.0 \times 10^{-5} \text{ T}$ south
- B $1.0 \times 10^{-4} \text{ T}$ north
- C $1.0 \times 10^{-4} \text{ T}$ south

The following information applies to questions 8–10.

The following diagram shows a conductor of length l carrying a current, I , perpendicular to a magnetic field of strength B . Ignore the magnetic field of the Earth.

- 8 Calculate the magnitude and direction of the magnetic force on this conductor for the following sets of data:

- a $B = 1.0 \text{ mT}$, $l = 5.0 \text{ mm}$, $I = 1.0 \text{ mA}$
- b $B = 0.10 \text{ T}$, $l = 1.0 \text{ cm}$, $I = 2.0 \text{ A}$



c $B = 1.0 \text{ T}$, $l = 10 \text{ mm}$, $I = 5.0 \text{ A}$

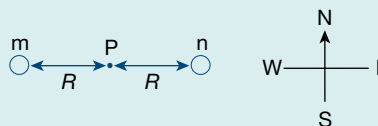
- 9 The direction of the magnetic field and current are changed. Describe the direction of the magnetic force on this conductor for:

- a B west, I north
- b B west, I south

- 10 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 11–14.

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 1.0 T at point P.



- 11 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

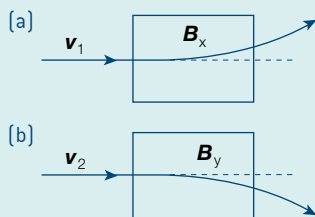
- 12 A third conductor, carrying a current, I , into the page, is placed at point P. What is the magnitude of the force per metre on this conductor if m and n both carry currents, I , into the page?

- 13 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 $I = 1.0 \text{ A}$?

- 14 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

- 15 The following diagrams (a) and (b) show two different electron beams being bent as they pass through two different regions of uniform magnetic field of equal magnitudes B_x and B_y . The initial velocities of the electrons in the respective beams are \mathbf{v}_1 and \mathbf{v}_2 .

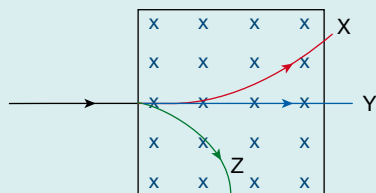


Which of the following is correct?

- A $v_1 = v_2$ B_x into the page
- B $v_1 = v_2$ B_y into the page
- C $v_2 > v_1$ B_y into the page
- D $v_2 < v_1$ B_y out of the page

16 Justify your answer to Question 15.

17 The following diagram shows the paths taken by three different particles, X, Y and Z, as they pass through a uniform magnetic field.

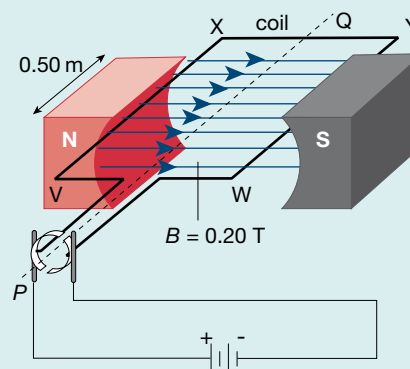


Which of the following could be correct?

- 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.

The following information applies to questions 18–20.

The diagram shows a simplified version of a direct current motor.



18 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.

19 In which direction will the coil begin to rotate?

20 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

Electromagnetic induction

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 electromagnetic induction, including:

- the generation of an EMF by electromagnetic induction
- the factors which 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.

10.1 Magnetic flux and induced currents

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.



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 field, \mathbf{B} , and the area A_\perp . A_\perp is the effective area of the loop perpendicular to the field lines. (If the loop is not perpendicular to the field, A_\perp will equal $A\cos\theta$, where θ is the angle between the field and the *normal* to the coil.)



MAGNETIC FLUX is given by:

$$\Phi_B = BA_\perp = BA\cos\theta$$



Figure 10.1 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.

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.

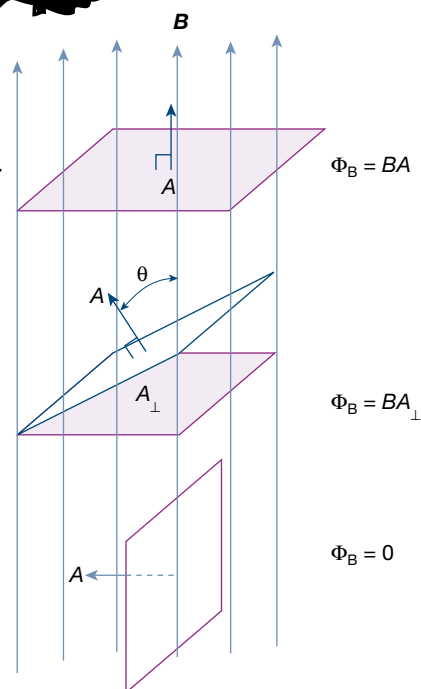


Figure 10.3 The magnetic flux is the strength of the magnetic field, \mathbf{B} , multiplied by the effective area ($A_{\perp} = A \cos \theta$), here shown shaded.

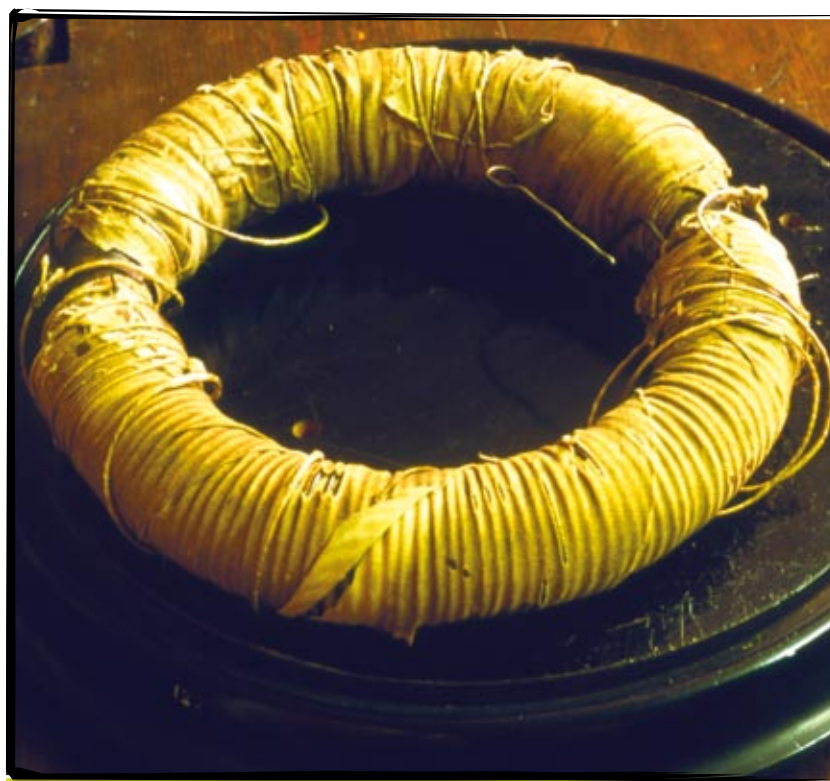


Figure 10.2 The original apparatus with which Faraday discovered electromagnetic induction.

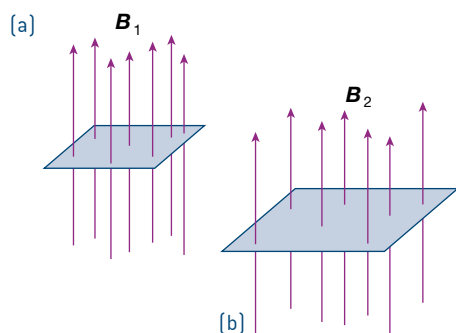


Figure 10.4 (a) A strong field. (b) A weaker field, but the same amount of flux.

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 common expression for the strength of the magnetic field, \mathbf{B} , as the *magnetic flux density*. \mathbf{B} can be thought of as proportional to the number of lines of force passing through a unit of area perpendicular to the lines.

Thus, the unit of magnetic flux is the magnetic field (tesla) multiplied by area (m^2), or T m^2 . This unit is also known as the weber ($1 \text{ Wb} = 1 \text{ T m}^2$). Conversely, this implies that the unit for magnetic field can be expressed as the flux density, or 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 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:

$$I \propto \frac{\Delta \Phi_B}{\Delta t}$$

Worked example 10.1A

A student places a horizontal 5 cm square coil of wire into a uniform vertical magnetic field of 0.1 T.

- How much flux cuts through the coil?
- She then pulls the coil out of the field at such a rate that it takes 1 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 mA is registered on the ammeter connected to it. If she puts the coil back into the flux at the same rate, what current will she observe?
- 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 s to rotate through 180° . Describe the current that will be induced.

Solution

- $\Phi_B = BA_\perp$
 $= 0.1 \times 0.05^2$
 $= 2.5 \times 10^{-4} \text{ Wb} \text{ (} 1 \text{ T m}^2 = 1 \text{ Wb)}$
- The same current but in the opposite direction: -2 mA .
- $I_{\text{in}} \propto \Delta\Phi_B/\Delta t$. This time Δt was halved so the current will be doubled. The current will be $+4 \text{ mA}$.
- The total change of flux that occurs in the 2 s is $-5 \times 10^{-4} \text{ Wb}$ (as the flux changes from $+2.5 \times 10^{-4}$ to $-2.5 \times 10^{-4} \text{ Wb}$). The average rate of this flux change is the same as in part a [$-2.5 \times 10^{-4} \text{ Wb s}^{-1}$], so the average current will again be $>2 \text{ mA}$. However, this will not be uniform. It will be a maximum ($>2 \text{ mA}$) as the loop passes through the vertical. (Although the flux is momentarily zero when the coil is vertical, this is the point of maximum flux change.)



10.1 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:
$$\Phi_B = BA_\perp = BA\cos\theta$$
- 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.

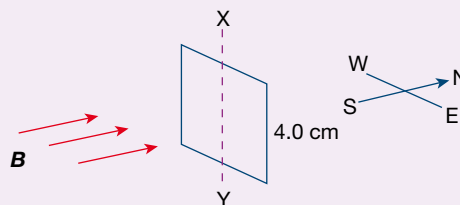


10.1 questions

Magnetic flux and induced currents

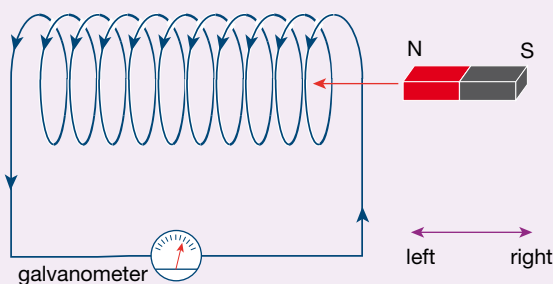
The following information applies to questions 1–3.

A square loop of wire, of side 4.0 cm, is in a region of uniform magnetic field, $B = 2.0 \times 10^{-3} \text{ T}$ north, as in the following diagram. The loop is free to rotate 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 a normal to the plane of the loop and north is 0° .



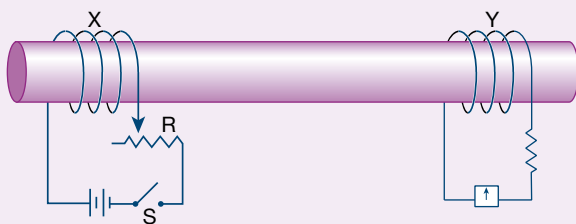
- 1 What is the magnetic flux passing through the loop when θ has the following values 0° , 45° , 60° , 90° ?
- 2 Calculate the magnitude of the change in magnetic flux through the loop when θ changes from 0° to 45° , 0° to 60° , 45° to 90° , 0° to 90° .
- 3 The loop is fixed in its initial position ($\theta = 0^\circ$). 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.

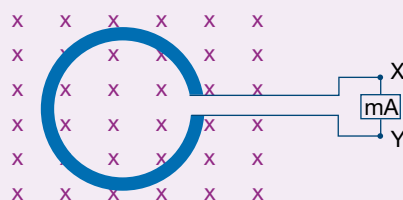


The switch, S, is initially open. When S is closed, a positive current flows through coil X.

- 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 4.0 cm has its plane perpendicular to a uniform magnetic field of strength 2.0 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 ms to drop to zero. When switched on again, it also takes 1 ms. Switching off the field results in a momentary current of 4.0 mA flowing through the millimeter 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 1.0 ms.
 - b The original coil is replaced with one of radius 2.0 cm and the original field is then switched off.
 - c The original field is switched off but it takes 2.0 ms instead of 1.0 ms to decrease uniformly to zero.

10.2 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 charges moving in a magnetic field 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 a charge q moves through a perpendicular magnetic field \mathbf{B} at a speed v , there is a magnetic force on it equal to qvB . The right-hand rule tells us that the force on positive charges in the wire in Figure 10.5 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 qvB force will do work on the electron as it moves along length l (the length of the wire in the field) given by:

$$W = Fl = qvBl$$

Given that there are many other charges being moved as well, this work will go into the potential energy because of 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:

$$\text{Potential difference } (\Delta V) = \frac{W}{q} = \frac{qvBl}{q} = vBl$$

This then is the EMF (\mathcal{E}) generated as a wire moves through a field. So for a single wire of length l moving at speed v in a perpendicular field \mathbf{B} the EMF is given by $\mathcal{E} = vBl$.

Worked example 10.2A

As an aeroplane flies along through the Earth's magnetic field, it is acting rather like the straight wire described above.

- At which places on the Earth will the induced EMF across an aeroplane's wings be maximum?
- What is the maximum EMF which would be induced across the wings of a Boeing 747 with a wing span of 64 m, flying at 990 km h^{-1} through the Earth's magnetic field? Take the maximum magnitude of the Earth's magnetic field as $60 \mu\text{T}$.

Solution

- 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.
- The maximum magnitude of the induced EMF near the poles is found from $\mathcal{E} = vBl$.

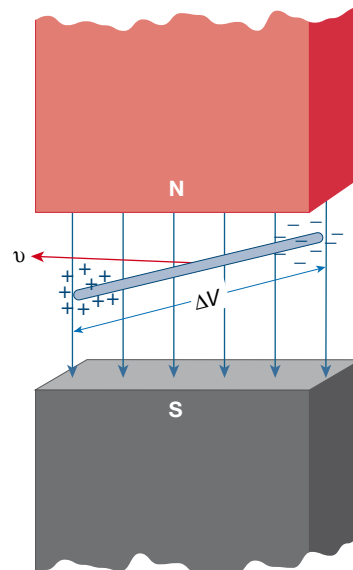


Figure 10.5 A potential difference ΔV will be produced across a straight wire moving to the left in a downward magnetic field.

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.

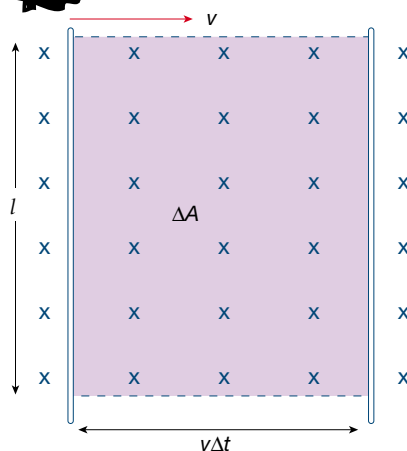


Figure 10.6 The area ΔA [shaded] swept out by the wire in the field in a time Δt is equal to $lv\Delta t$.

Physics file

The EMF, or electromotive force (\mathcal{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.

$$v = 990 \text{ km h}^{-1}$$

$$= \frac{990}{3.6}$$

$$= 275 \text{ m s}^{-1}$$

$$B = 60 \mu\text{T}$$

$$= 60 \times 10^{-6} \text{ T}$$

$$= 6 \times 10^{-5} \text{ T}$$

So:

$$\mathcal{E} = vBl$$

$$= 275 \times 6 \times 10^{-5} \times 64$$

$$= 1.06 \text{ V}$$

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 $\mathcal{E} = vBl$ 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 $v\Delta t$, so the area ΔA it has swept out is then equal to $lv\Delta t$. This means that the amount of magnetic flux swept through by the wire in this time is given by:

$$\Delta\Phi_B = B\Delta A = Blv\Delta t$$

This can be written

$$Blv = \frac{\Delta\Phi_B}{\Delta t}$$

But the EMF, \mathcal{E} , was just equal to vBl , so

$$\mathcal{E} = \frac{\Delta\Phi_B}{\Delta t}$$

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.



FARADAY'S LAW OF INDUCTION states that:

$$\mathcal{E} = -\frac{\Delta\Phi_B}{\Delta t}$$

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.

Notice that a negative sign has appeared in the expression. This has an important significance (which will be explained in section 10.3), 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:

$$\mathcal{E} = -N \frac{\Delta\Phi_B}{\Delta t}$$

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 \mathcal{E}/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 10.2B

A student winds a coil of area 40 cm^2 with 20 turns. He places it horizontally in a vertical uniform magnetic field of 0.1 T , and connects it to a galvanometer with resistance 200Ω .

- How much flux passes through the coil?
- If it is then withdrawn from the field in a time of 0.5 s , what would be the average current reading on the galvanometer?

Solution

- $\Phi_B = BA_{\perp}$. Here $B = 0.1 \text{ T}$ and $A_{\perp} = 40 \text{ cm}^2 = 0.0040 \text{ m}^2$. Thus the flux through the coil is $\Phi_B = 0.1 \times 0.0040 = 4 \times 10^{-4} \text{ Wb}$.

- The average EMF induced in each turn is given by the average rate of change of flux, i.e.

$$\begin{aligned}\mathcal{E} &= \frac{\Delta\Phi_B}{\Delta\tau} \\ &= \frac{4 \times 10^{-4}}{0.5} \\ &= 8 \times 10^{-4} \text{ V}\end{aligned}$$

Hence the total EMF generated in the coil of 20 turns will be:

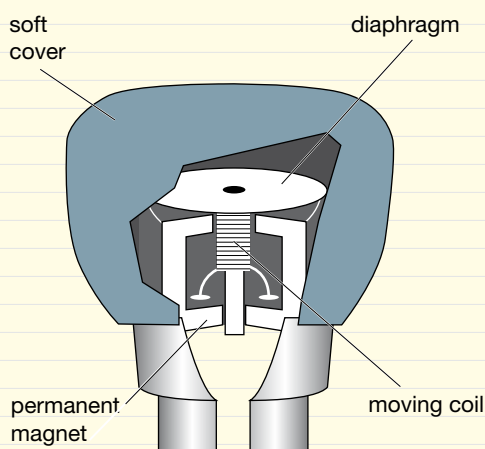
$$\begin{aligned}\mathcal{E}_{\text{coil}} &= 20 \times 8 \times 10^{-4} \\ &= 0.016 \text{ V}\end{aligned}$$

Assuming that the only significant resistance in the circuit is the galvanometer, the average induced current is given by Ohm's law:

$$\begin{aligned}I &= \frac{V}{R_t} \\ &= \frac{0.016}{200} \\ &= 8 \times 10^{-5} \text{ A or } 80 \mu\text{A}\end{aligned}$$

Physics in action

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.

Figure 10.7 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.

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.

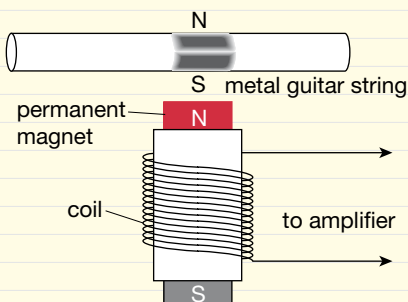


Figure 10.8 A guitar pick-up magnetises the steel string with a permanent magnet and then uses this vibrating magnet to induce an EMF in a coil below it.



10.2 summary

Induced EMF: Faraday's law

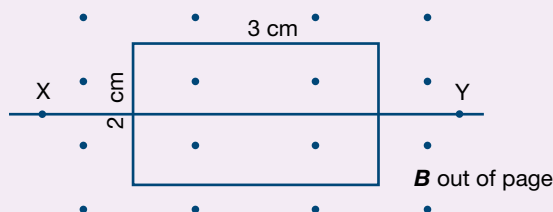
- Faraday found that the EMF induced in a loop was dependent only on the rate of change of magnetic flux through the loop.
- As a conductor moves perpendicular to a magnetic field, work is done on the moving charges to produce a potential difference along the wire.
- This potential difference appears 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.



10.2 questions

Induced EMF: Faraday's law

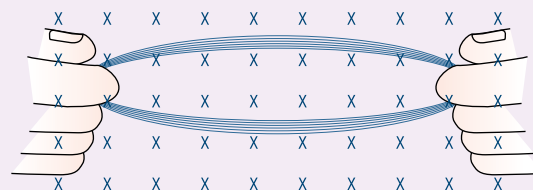
- A rectangular wire loop is located with its plane perpendicular to a uniform magnetic field of 2.0 mT , directed out of the page, as shown. The loop is free to rotate about a horizontal axis XY and has a resistance of 1.5Ω .



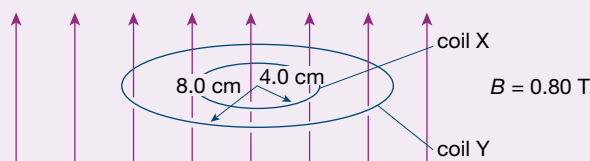
- How much magnetic flux is threading the loop in this position?
 - The loop is rotated about XY , through an angle of 90° , so that its plane becomes parallel to the magnetic field. How much flux is threading the loop in this new position?
 - Calculate the average EMF induced in the loop if this rotation took 40 ms .
 - What is the average current induced in the loop?
- A coil of 500 turns, each of area 10 cm^2 , is wound around a square frame. The plane of the coil is

initially parallel to a uniform magnetic field of 80 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 ms .

- What is the change in magnetic flux through each turn of the coil during this time?
 - What is the average EMF induced in each turn during the rotation?
 - Calculate the average EMF induced in the coil during this time.
- A student stretches a flexible wire coil of 30 turns and places it in a uniform magnetic field of strength 5.0 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 cm^2 to 250 cm^2 in 0.50 s .



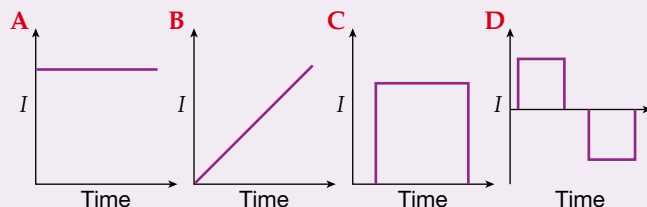
- a What is the change in magnetic flux through each turn of the coil during this time?
- b Find the average EMF induced in the coil during this time.
- 4 The vertical component of the Earth's magnetic field in Melbourne is 5.0×10^{-5} T upwards. A helicopter at Melbourne Airport has metal rotor blades 4.0 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 8.0 revolutions per second, calculate the average EMF induced between the axis and tip.
- 5 Two flat, circular coils, X of radius 4.0 cm and 20 turns and Y of radius 8.0 cm and 10 turns, are located concentrically in a horizontal plane perpendicular to a uniform magnetic field of strength 0.80 T, as shown. If the direction of the magnetic field is reversed in 0.10 s, calculate the ratio of the average EMF induced in coil X to that induced in coil Y.



The following information applies to questions 6 and 7. 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.

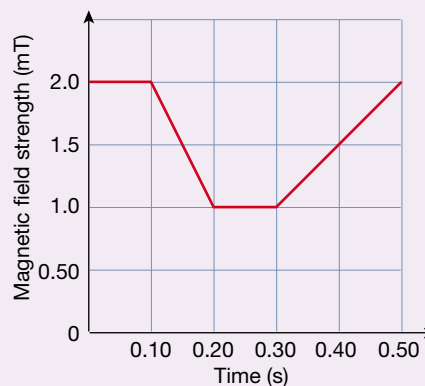


- 6 Which of the graphs A–D best represents the current I in the loop as a function of time t ?



- 7 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.

- 8 During a physics experiment, a student pulls a rectangular conducting loop from between the poles of a magnet in 0.10 s. The loop has an area of 8.0 cm² and contains 10 turns. Initially the entire loop is located in a region of uniform magnetic field of strength 0.10 T. The final position is free of the field.
- a What current flows through the loop in its initial position?
- b Calculate the average EMF induced in the loop by the student's action.
- c Without changing the coil or magnet, how could the student generate an EMF of greater magnitude?
- 9 A conducting loop containing 100 turns, each of area 2.0×10^{-3} m², has its plane perpendicular to a magnetic field whose strength varies with time according to the following graph.



What is the magnitude of the average EMF induced in the loop:

- a between $t = 0$ and $t = 0.10$ s?
- b between $t = 0.10$ s and $t = 0.20$ s?
- c between $t = 0.20$ s and $t = 0.30$ s?
- d between $t = 0.30$ s and $t = 0.50$ s?
- 10 A record collector plays an Elvis LP, made of platinum, on a turntable that rotates at 33 revolutions per minute. Assume that the direction of the Earth's magnetic field is always perpendicular to the plane of the record, and has a uniform strength of 5.0×10^{-5} T. The radius of the record is 26 cm. What EMF would be induced between the centre of the record and the rim? Would an EMF be induced between opposite edges of the record?

10.3

Direction of EMF: Lenz's law

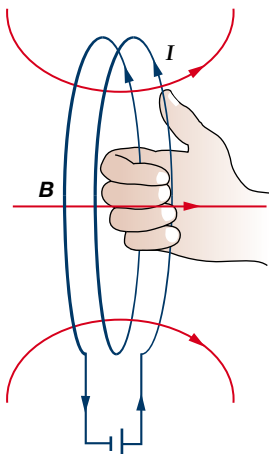


Figure 10.9 Using the right-hand grip rule to find the direction of flux from the current in a coil, or vice versa.

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.

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 10.10a. 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 10.9) 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 10.10b)—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.

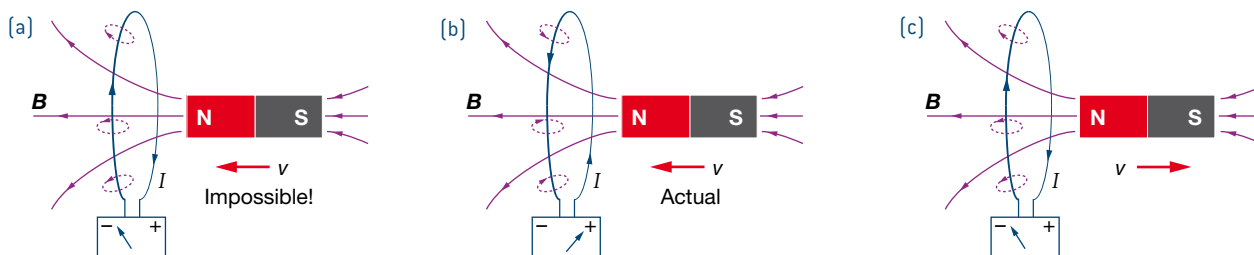


Figure 10.10 (a) An impossible situation! As the magnet approaches, the induced current would rapidly escalate, contravening the principle of conservation of energy. (b) The actual induced current opposes the changing flux. (c) As the magnet is withdrawn, the induced current creates a flux, which opposes 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 10.10c). 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.

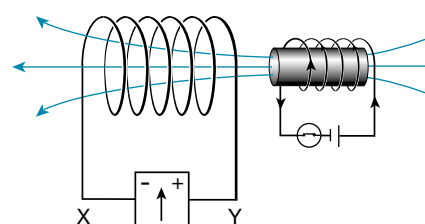
Worked example 10.3A

Instead of using a permanent magnet to change the flux in the loop in the example just given, 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 above. 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 *IIB* 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

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 10.11 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.



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 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 10.12). 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



Figure 10.12 A disk magnet is repelled by a superconductor because the magnet induces a permanent current into the superconductor, which results in an opposing field.

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.



10.3 summary

Direction of EMF: Lenz's law

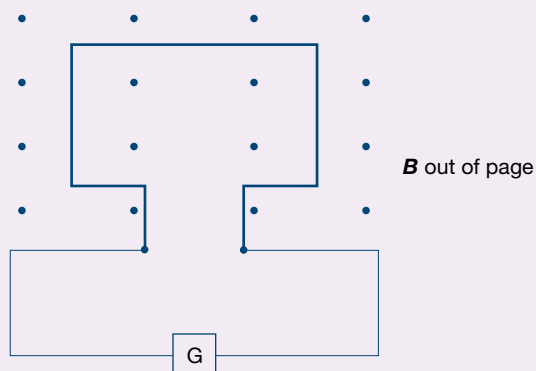
- Lenz's law states that induced current in a loop will be in the direction so that the flux it creates will oppose the change in the flux that produced it.
- Lenz's law is an expression of the principle of conservation of energy.
- Induced currents set up by the relative motion of a conductor and magnet will create a field that will apply a force that will oppose the relative motion.



10.3 questions

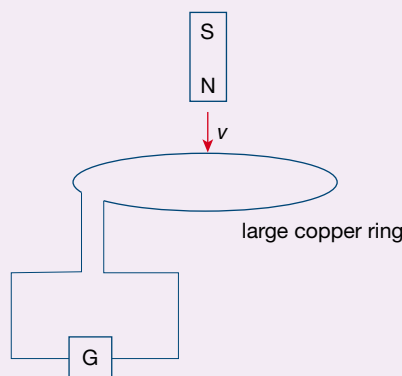
Direction of EMF: Lenz's law

- 1 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.
- 2 A square 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.

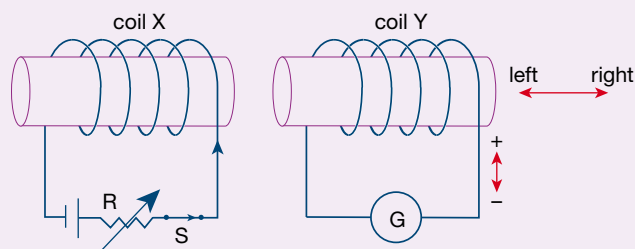


What is the direction of the magnetic field due to the induced current when the following changes are made to the initial field?

- a The field is switched off.
 - b The field strength is doubled.
 - c The direction of the field is reversed.
- 3 A bar magnet is falling towards the centre of a horizontal copper ring, as shown.



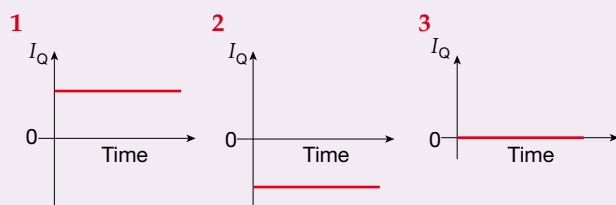
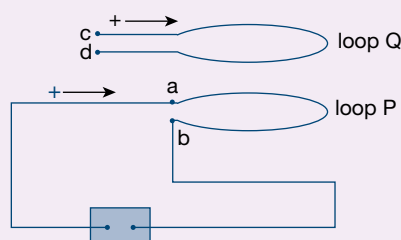
- a What is the direction (clockwise or anticlockwise as seen from above) of the induced current in the ring when the magnet is in the position shown in the diagram?
 - b Describe the direction of the induced current in the ring just after the magnet has passed completely through the ring.
 - c At what instant does the direction of the induced current in the ring change?
 - d Name four factors that would influence the magnitude of the induced current in the copper ring.
- 4 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.

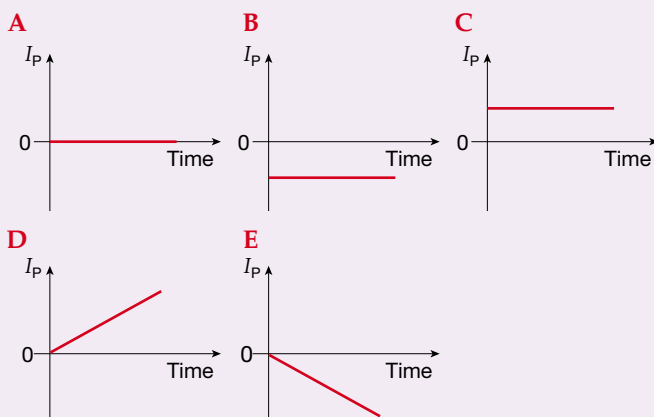
- a S is opened.
- b R is increased.
- c R is reduced.

- 5 Two conducting loops P and Q are located with their planes parallel to each other as shown below. An electric current, I_P , can flow in either direction through loop P, and the induced current, I_Q , in loop Q can be measured. Positive I_P is from a to b through loop P. Positive I_Q is from c to d through loop Q. An experiment is performed in which the current through P is varied with respect to time in three different ways. The resulting graphs 1–3 of the induced current I_Q in Q versus time are shown.



Which one or more of the following graphs (A–E) represents the I_P versus time relationship which could have caused I_Q to behave in the way shown in:

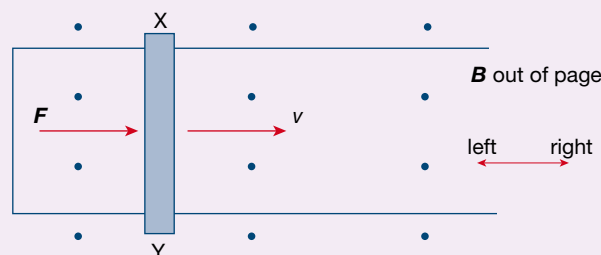
- a graph 1? b graph 2? c graph 3?



Justify your answers.

The following information applies to questions 6–8.

A metal rod XY is being pushed along part of a copper rectangle by an external force, F , as shown. The velocity, v , of the rod is perpendicular to a uniform magnetic field directed out of the page.



- 6 a What is the direction of the induced current in the rod?
- b What is the direction of the magnetic force on the rod due to the current induced in it?
- 7 The rod is still being pushed along the rails to the right, but the direction of the magnetic field is reversed, i.e. it is now directed into the page.
 - a What is the direction of the induced current in the rod now?
 - b What is the direction of the magnetic force on XY now?
- 8 Which of the following is correct? The induced current in the rod will always produce a force whose direction:
 - A is opposite to the direction of the external force
 - B is the same as the direction of the external force
 - C depends on the direction of the external magnetic field.

10.4 Electric power generation

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 Melbourne and Sydney 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 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.

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

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 10.6: 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.

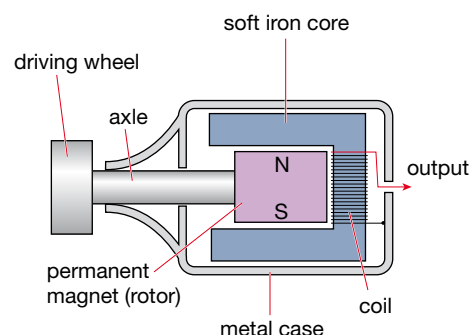


Figure 10.13 A bicycle dynamo is a simple rotating magnet generator. Because the current is taken from fixed coils there are few moving parts and little friction. The 'magneto' which powers the spark plug in small two-stroke engines uses a similar principle.

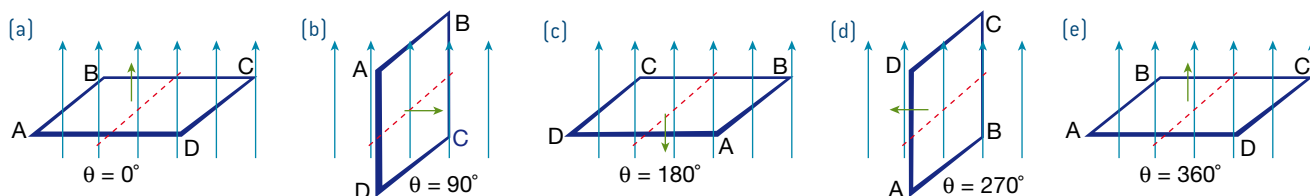


Figure 10.14 (a) The field \mathbf{B} and the plane of the area A of the loop are perpendicular and the amount of flux through the loop is maximum $\Phi_B = BA$. (b) After the loop has turned through a quarter of a revolution the plane of the loop is parallel to the field and the flux through the loop is zero. (c) The flux then increases but in the opposite sense relative to the loop; $\Phi_B = -BA$. 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.

Physics file

Lenz's law tells us that the induced current creates a field to oppose the *change* of flux. 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.

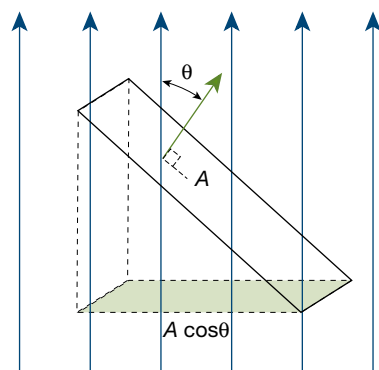


Figure 10.15 The angle θ is the angle between the field and the normal to the area. It is usual to describe the orientation of a plane by the direction of the normal, as otherwise two directions would be required.

Physics file

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: $50 \times 2\pi = 100\pi \text{ rad s}^{-1}$ (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:

$$\mathcal{E} = \frac{d\Phi_B}{dt} = -\frac{d(BA \cos \omega t)}{dt} = BA\omega \sin \omega t$$

coil increases and decreases. Consider the simple case of a rotating loop in a constant, uniform magnetic field, as shown in Figure 10.14.

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 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. 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 through the loop is given by $\Phi_B = BA \cos \theta$, where θ is the angle between the field (\mathbf{B}) and the normal to the plane of the area (\mathbf{A}).

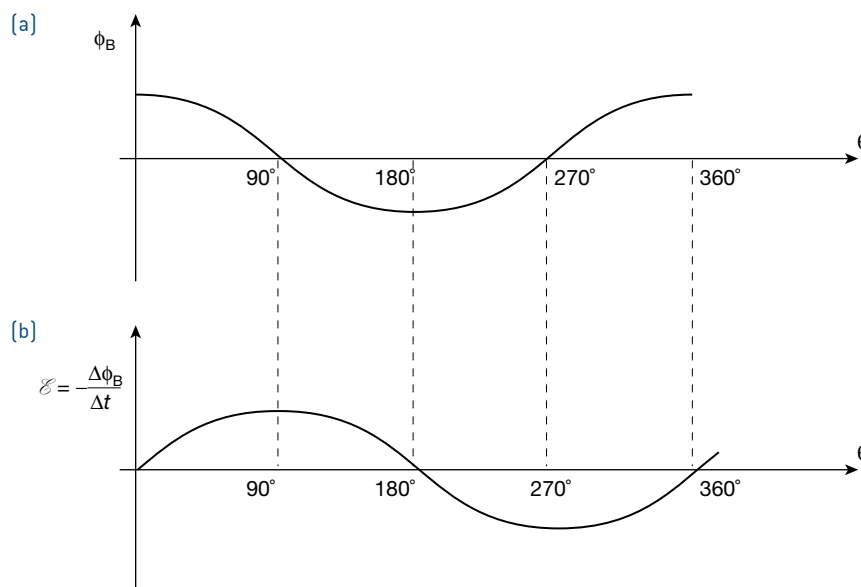


Figure 10.16 (a) The flux Φ_B through the loop as a function of the angle θ between the field and the normal to the loop. (b) The rate of change of flux $\Delta\Phi_B/\Delta t$ through the loop (and hence the EMF, \mathcal{E}) as a function of the angle θ between the field and the normal to the loop. The loop is rotating at a steady speed.

As we can see from Figure 10.16, 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 is not hard but requires a little calculus. We have derived it in the adjacent Physics file. That expression is for a single turn loop. For a coil of N turns the EMF will be N times as much.



For a coil of area A and N turns, rotating in a magnetic field \mathbf{B} at an angular speed ω , the EMF generated is given by:

$$\mathcal{E} = NBA\omega \sin \omega t$$

Notice that the maximum (peak) value of the EMF is given by $NBA\omega$. As we could well expect, it is proportional to the *strength of the field*, the *area* and *number of turns* of the coil, and the *rate* at which it is turning.



INTERACTIVE TUTORIAL
AC generators

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 9. In the first case, the output will be just the alternating current which results from the EMF shown in Figure 10.17a, 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 10.17b). The DC generator was widely used in cars up until the 1980s when semiconductor diodes enabled the simple conversion of AC into DC.

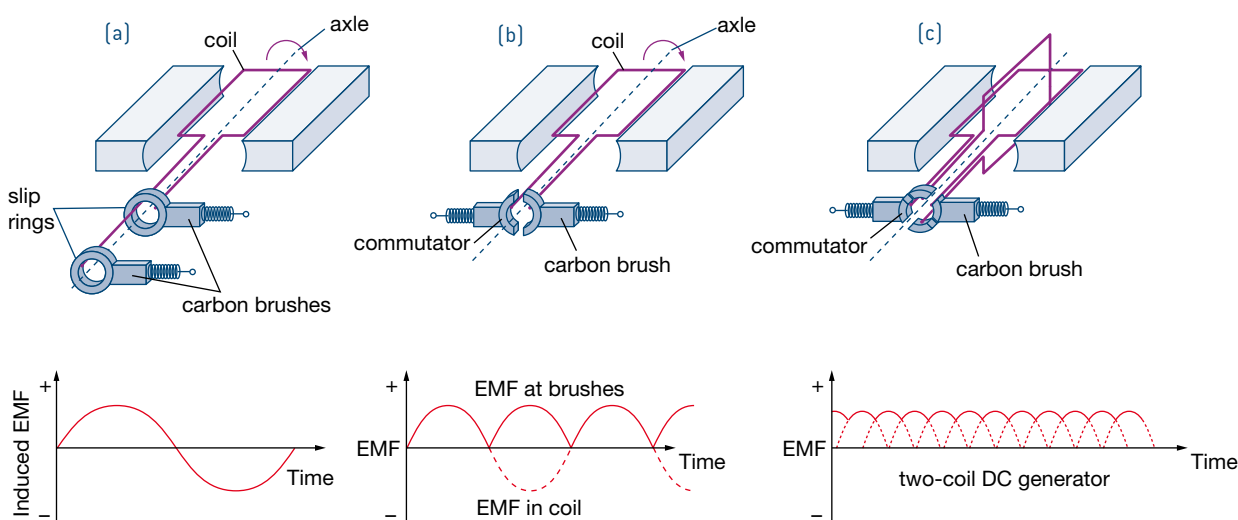


Figure 10.17 (a) An AC generator uses simple slip rings to take the current from the coil. (b) A DC generator has a commutator to reverse the direction of the alternating current every half cycle and so produce a DC output. (c) Two (or more) coils may be used to smooth the DC output.

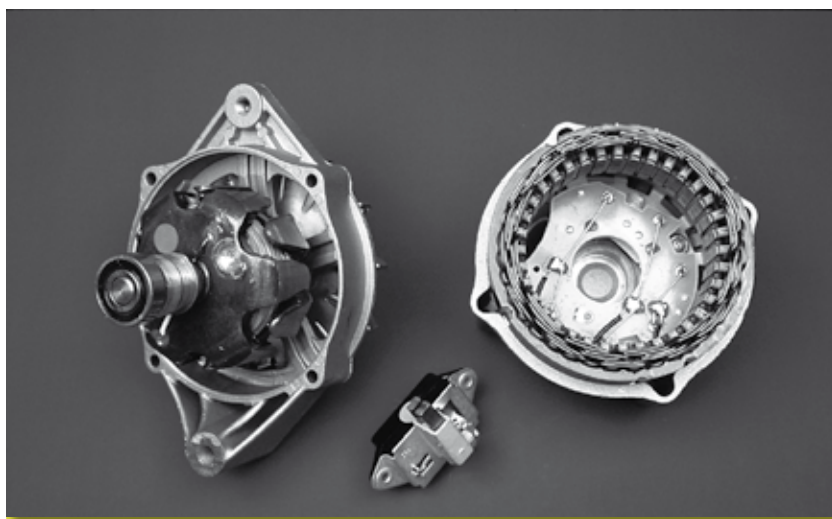


Figure 10.18 A modern car alternator employs a rotating electromagnet which induces an alternating EMF in the stator coils. This AC is often rectified by diodes inside the housing and so the output from the device is actually DC.

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.

Power station generators

The large generators used in power stations are rotating magnet alternators. The DC current used for the rotor of a typical generator is around 2500 A. The output voltage may be 20 000 V with a current of 17 500 A, giving a power output of 350 MW. The alternating current is fed, via transformers, straight into the electricity grid.

The energy used to drive the generators in Victoria's Latrobe Valley is obtained from burning brown coal. The heat from the coal is used to generate steam at temperatures of over 500°C and pressures of around 160 atm. This steam is used to drive three stages of turbines. After the steam goes through the first stage it is reheated and sent through the second and third stages, each stage operating at a lower pressure and temperature than the one before. This enables energy to be extracted from the steam and converted into mechanical energy in the most efficient way. In fact, modern turbines can operate at efficiencies of up to 40%. (Compare this with a typical car engine, which has an efficiency of only about 15%.)

Victoria's coal-fired power stations operate at an overall efficiency of about 26% because the brown coal they use has a considerable moisture content; it takes about 23% of the energy content to evaporate the water. The other 51% is wasted as heat. The generator itself converts the mechanical energy of the rotating turbine into electrical energy at around 98% efficiency.



Figure 10.19 The turbine of a steam-driven generator.

The result of this unavoidable inefficiency is that large amounts of heat are wasted in the process of converting the steam energy into mechanical energy. Most of it goes into the water used to condense the steam once it has done its work.



Figure 10.20 The steam is from the cooling water used in the heat exchangers (condensers) to condense the steam after it has passed through the turbines. The water used in the boilers and turbines is extremely pure and is reheated after being condensed.



10.4 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.

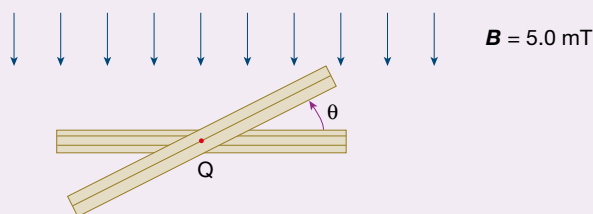


10.4 questions

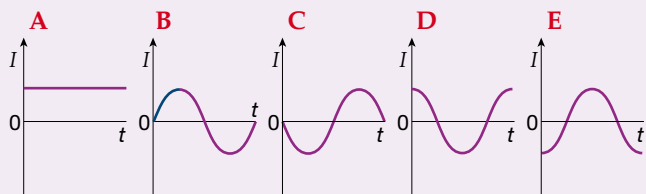
Electric power generation

The following information applies to questions 1–6.

The following diagram shows a square coil containing 100 turns each of area 20 cm^2 , mounted on a horizontal axis through Q. The plane of the coil is initially perpendicular to a vertical uniform magnetic field of 5.0 mT . The coil is to be rotated about Q in a counter-clockwise direction at a rate of 15° per millisecond (i.e. a frequency of 42 Hz).

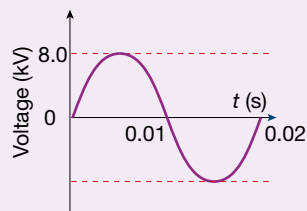


- What is the magnetic flux (Φ_B), in mWb, at each of the following values of θ ?
a 0° **b** 15° **c** 30° **d** 45°
e 60° **f** 75° **g** 90°
- Calculate the average rate of change of flux through the coil when it is rotated through each of the six 15° intervals from 0° to 90° as in Question 1.
- Considering your answers to Question 2, what do you notice about the rate of change of flux through the coil as it rotates from $\theta = 0^\circ$ to $\theta = 90^\circ$?
- Determine the average EMF induced in the coil during each of the first six 1.0 ms intervals of its rotation.
- 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?
- Assuming that a counter-clockwise rotation of the coil from $\theta = 0^\circ$ initially produces a positive current, which of the graphs best illustrates the variation of the induced current as a function of time?

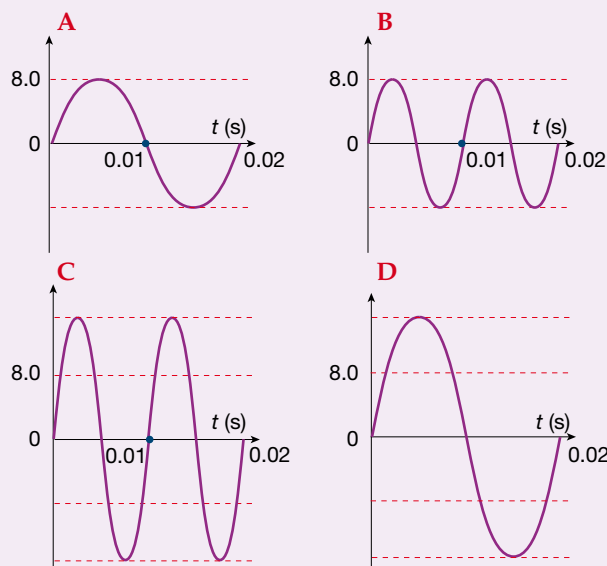


The following information applies to questions 7 and 8.

A simple generator consists of a coil of $N = 1000$ turns each of radius 10 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 Hz .



- Which of the graphs below best describes the voltage–time relationship of this generator for the following modifications?
 - $N = 1000$, $r = 10 \text{ cm}$, field $= B$, $f = 100 \text{ Hz}$
 - $N = 1000$, $r = 10 \text{ cm}$, field $= 2B$, $f = 50 \text{ Hz}$
 - $N = 500$, $r = 10 \text{ cm}$, field $= 2B$, $f = 100 \text{ Hz}$
 - $N = 1000$, $r = 5.0 \text{ cm}$, field $= 2B$, $f = 100 \text{ Hz}$
 - $N = 2000$, $r = 10 \text{ cm}$, field $= B$, $f = 50 \text{ Hz}$



- Calculate the strength of the magnetic field required to produce a peak voltage of 8.0 kV .

10.5 Alternating voltage and current

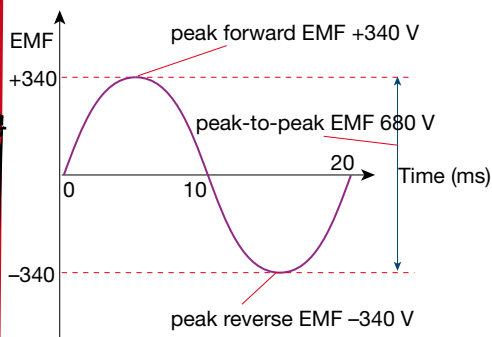


Figure 10.21 The voltage in our power points oscillates between +340 V and -340 V 50 times each second. The value of a DC supply that would supply the same average power is 240 V.

Physics file

In an AC circuit, the power produced in a resistor is equal to $(V_p^2/R)\sin^2\theta$. As can be seen from Figure 10.22, the *average power* will be given by $\frac{1}{2}V_p^2/R$. 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:

$$V_{av}^2/R = \frac{1}{2}V_p^2/R$$

which simplifies to:

$$V_{av}^2 = \frac{1}{2}V_p^2$$

$$\text{or } V_{av} = \frac{V_p}{\sqrt{2}}$$

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.

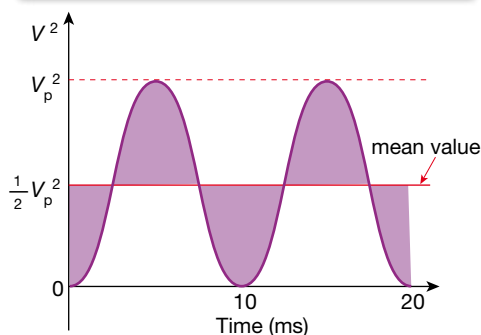


Figure 10.22 The square of the voltage for an AC supply. The average value of V^2 is equal to $\frac{1}{2}V_p^2$. Note that the frequency is 100 Hz.

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 times 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. $V_{p-p} = 680$ 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 don't respond to flickering at frequencies greater than about 20 Hz.

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 $\sqrt{2}$. This is referred to as the *root mean square* or RMS voltage.

$$V_{RMS} = \frac{V_p}{\sqrt{2}} \text{ or } V_p = \sqrt{2} V_{RMS}$$

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: $I_{RMS} = I_p/\sqrt{2}$ or $I_p = I_{RMS}\sqrt{2}$.

The mains voltage supplied to our houses has a peak value of 340 V. It is the RMS value of the voltage, $340/\sqrt{2} = 240$ 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.

Worked example 10.5A

A 60 W light bulb supplied with 240 V AC uses 60 J every second, but the instantaneous power varies from 0 to 120 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 W is the RMS power. It is obtained as the product of the RMS voltage and RMS current. The RMS voltage is 240 V and so the RMS current is given by:

$$\begin{aligned} I_{\text{RMS}} &= \frac{P}{V_{\text{RMS}}} \\ &= \frac{60}{240} \\ &= 0.25 \text{ A} \end{aligned}$$

The peak power is given by:

$$\begin{aligned} P_p &= I_p V_p \\ &= \sqrt{2} I_{\text{RMS}} \times \sqrt{2} V_{\text{RMS}} \\ &= 2 I_{\text{RMS}} V_{\text{RMS}} \\ &= 2 \times 0.25 \times 240 \\ &= 120 \text{ W} \end{aligned}$$



Figure 10.23 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.

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 10.24. 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 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 10.24b). At any time the average current is actually zero.

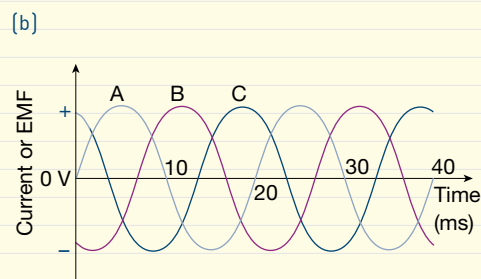
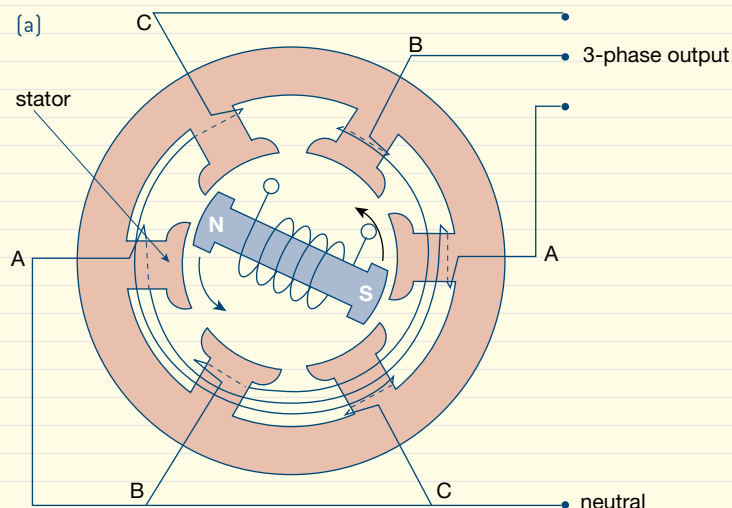


Figure 10.24 (a) A three-phase generator. The DC current is fed to the rotating magnet via slip rings (not shown). The ends of each of the three stator coils are connected together—this becomes the ‘neutral’. The other ends carry the high AC voltages which are one-third of a cycle apart, as shown in (b).



10.5 summary

Alternating voltage and current

- 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 $V_{p-p} = 680$ V.
- The root mean square voltage (V_{RMS}) is the value of an equivalent steady voltage (DC) supply which

would provide the same power. $V_{RMS} = V_p / \sqrt{2}$. The RMS value of domestic mains voltage is 240 V.

- For simple resistive circuits the current is given by Ohm's law, and $I_{RMS} = I_p / \sqrt{2}$.
- The average power in a resistive AC circuit is given by $P = V_{RMS} \times I_{RMS} = \frac{1}{2} V_p \times I_p$.



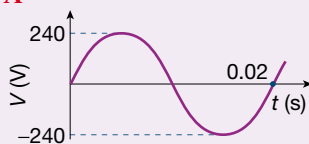
10.5 questions

Alternating voltage and current

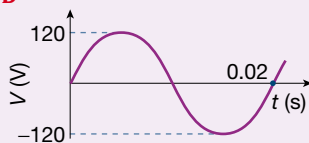
- Which of the following is true? A 240 V RMS voltage:
 - is the same as a 240 V DC voltage
 - is equivalent to a DC voltage of 340 V
 - is equivalent to a DC voltage of 170 V
 - will provide the same average power to a circuit as a 240 V DC voltage.

- Which graph best describes the variation of voltage with time for the household supply in Victoria?

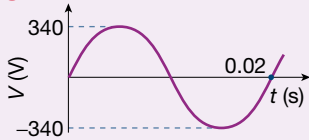
A



B



C



D

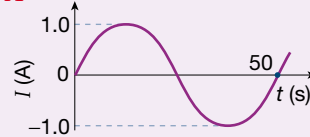


- An electrician is called to check a household power point and confirms that the voltage between the active (A) and neutral (N) is 240 V RMS.
 - What is the peak voltage between A and N?

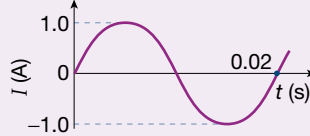
- What is the peak-to-peak voltage between A and N?
- An appliance of total resistance 100Ω is plugged into the socket. What is the peak current that will flow in the circuit?
- What is the RMS current that will flow in the circuit?

- An AC supply of frequency 50 Hz is connected to a circuit, resulting in an RMS current of 1.0 A being observed. Which graph best shows the variation of current with time for this circuit?

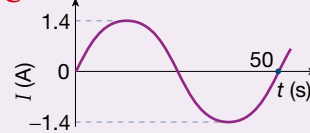
A



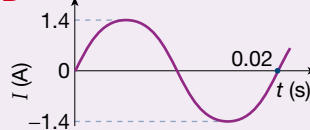
B



C



D



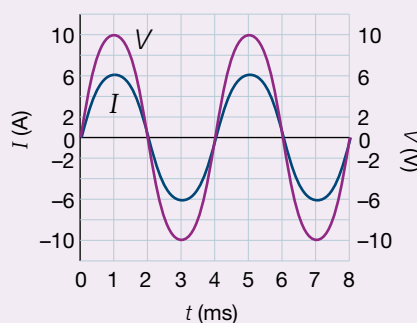
- The following diagram shows the voltage-time characteristics of a particular power supply.





- a What is the frequency of this power supply?
 - b What is the peak-to-peak voltage of the supply?
 - c What is the RMS voltage of the supply?
- 6 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?
 - 7 A loudspeaker is rated at 60 W RMS maximum at 24 V RMS.
 - a What is the RMS current in the speaker when it operates at maximum power?
 - b What is the peak current in the speaker when it is operating at its maximum output?
 - c What is the peak-to-peak voltage across the speaker when it is operating at maximum power?
 - d What is the apparent resistance of the speaker?
 - 8 A lamp purchased in the USA is designed to operate at optimum efficiency when it is connected to an AC supply that has a peak voltage of 170 V and a frequency of 60 Hz. The lamp has an operating resistance of 100 Ω .
 - a What DC voltage is required to operate this lamp at optimum efficiency?
 - b How much power would this lamp be consuming when operating at optimum efficiency on:
 - i AC?
 - ii DC?
 - c When this lamp is operating on AC, how many times per second will peak power occur?
 - d What is the peak power produced by this lamp?
 - 9 A student decides to test the output power of a new amplifier by using a CRO to display the alternating current I and voltage V that it produces. The result is shown.

 - a What is the RMS voltage output of the amplifier?
 - b What is the RMS current in the amplifier?
 - c What is the RMS power rating of the amplifier?



10.6 Transformers

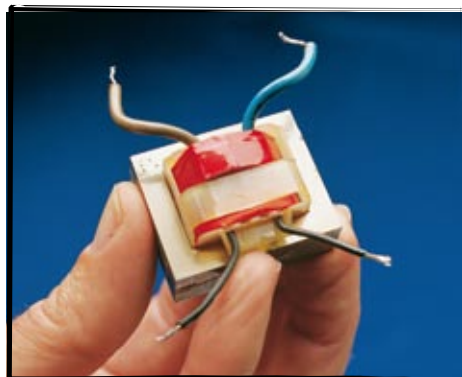


Figure 10.25 A modern transformer is most commonly made by winding the two coils around an iron core.

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 10.26 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.

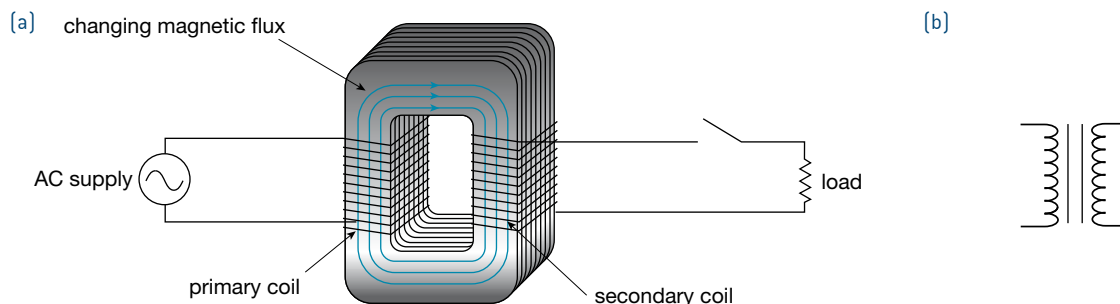


Figure 10.26 (a) In an ideal transformer, the iron core ensures that all the flux generated in the primary also passes through the secondary. (b) The symbol used in circuit diagrams for an iron core transformer.

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.

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 would 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:

$$\frac{V_s}{V_p} = \frac{N_s}{N_p}$$

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.

PRACTICAL ACTIVITY 35

Transformer operation

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!

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:

$$I_p V_p = I_s V_s$$

$$\text{or } \frac{I_s}{I_p} = \frac{V_p}{V_s}$$



The **TRANSFORMER EQUATIONS** are:

$$\frac{I_s}{I_p} = \frac{V_p}{V_s} = \frac{N_p}{N_s}$$

Worked example 10.6A

A transformer supplies a model train that operates on 12 V from the mains 240 V supply.

- If the number of turns in the secondary coil is 100, what will be the number of turns in the primary coil?
- If the train requires a current of up to 4 A, what will be the current and the power drawn from the mains?

Solution

- We know that the turns ratio $\frac{N_s}{N_p}$ is equal to the voltage ratio $\frac{V_s}{V_p}$.

Thus we can write:

$$\frac{N_s}{N_p} = \frac{V_s}{V_p} = \frac{240}{12} = 20$$

As there are 100 turns in the secondary there must be 2000 turns in the primary.

- The current ratio is the inverse of the turns and voltage ratio so that the current in the primary will be 1/20 of the secondary current. Thus:

$$I_p = \frac{4}{20} = 0.2 \text{ A}$$

The power required by the train was $12 \text{ V} \times 4 \text{ A} = 48 \text{ W}$. Alternatively, the input power was $240 \text{ V} \times 0.2 \text{ A} = 48 \text{ W}$, 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 ($P = I^2 R$), 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 10.27). 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 10.28. 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.

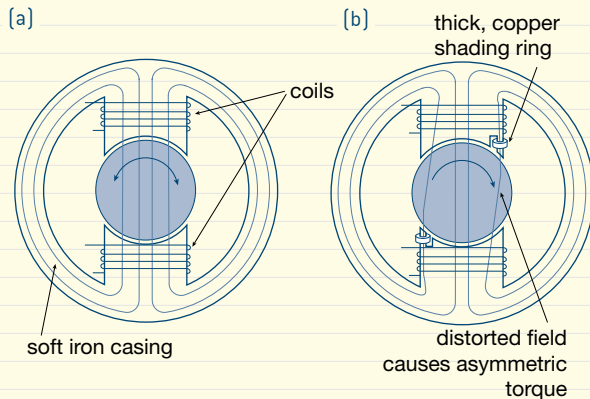


Figure 10.29 The principle of a simple single-phase, 'shaded pole' induction motor. The distorted (or 'shaded') field causes the rotor to turn in one direction in preference to the other.

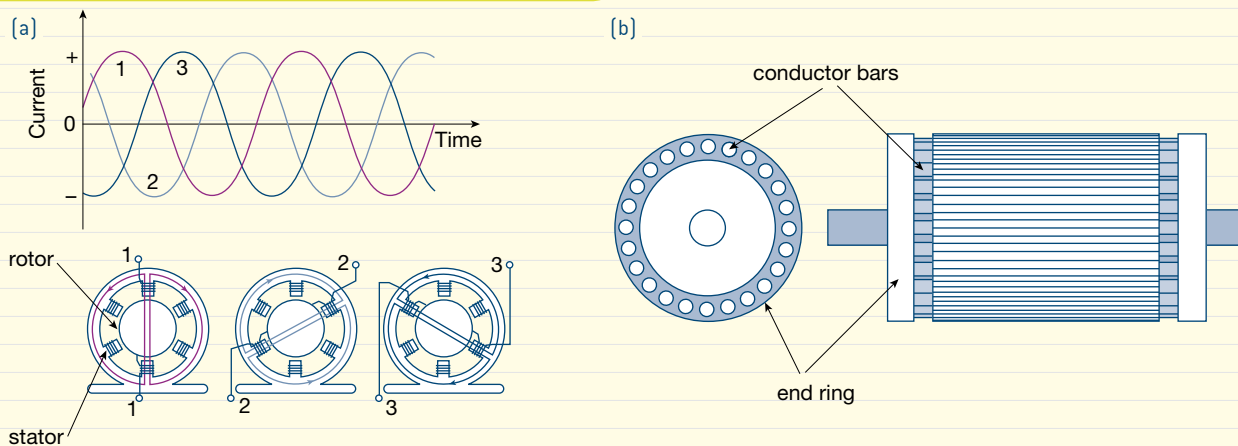


Figure 10.30 (a) In a three-phase motor, as the current in each pair of opposite coils peaks, the field appears to rotate, dragging the rotor around with it. (b) A 'squirrel cage' rotor. The rotor is made of iron laminations to cut down undesirable eddy currents. The induced currents flow lengthwise in copper or aluminium rods which are joined at the ends (as in a squirrel cage).

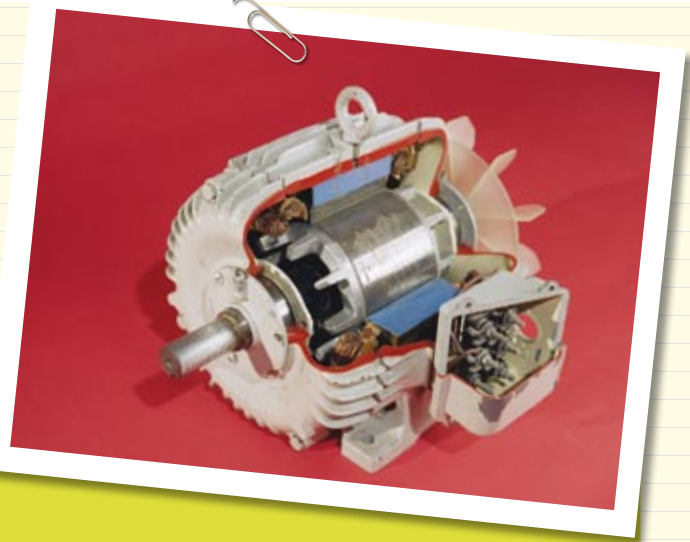


Figure 10.28 A typical AC induction motor. Induction motors are the most common electric motor used in industry. Because they have no brushes they are both more efficient and more reliable.



10.6 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:

$$V_s/V_p = N_s/N_p$$

- 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:

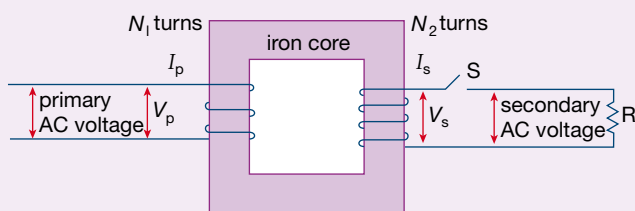
$$P_p = P_s, \text{ so } I_p V_p = I_s V_s \text{ or } I_p/I_s = V_s/V_p$$



10.6 questions

Transformers

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:

- a the input power to the primary coil of an ideal transformer?
 - b the 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:

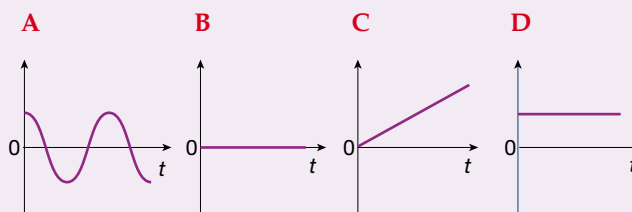
- a the power input to the primary coil of a non-ideal transformer?
 - b the power output from the secondary coil of a non-ideal transformer?
- A $V_p I_p$
 - B $V_s I_s$
 - C $V_p I_s$
 - D $I_s^2 R$

- 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 8.0 V and a primary RMS current of 2.0 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 V RMS and a RMS current of 2.0 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 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.

Physics file

$P = I^2R$ results from the substitution of Ohm's law $V = IR$ into the power equation $P = IV$.

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 $R = \rho l/A$, 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 ($\Omega \text{ m}$). Resistivity can be thought of as a measure of the inherent resistance of the material. The resistivity of copper is $1.7 \times 10^{-8} \Omega \text{ m}$ and that of aluminium is $2.8 \times 10^{-8} \Omega \text{ m}$.

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 der Graff does). The corona effect makes transmission voltages greater than this impracticable.

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.

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 Melbourne uses up to 8000 MW of power at peak times. At 250 V, this would require a current of $3.2 \times 10^7 \text{ A}$ (32 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 carried by a transmission line is the product of the current carried and the voltage ($P = IV$). The higher the voltage, the lower the current needed. At a voltage of 500 kV, the 8000 MW could be carried by a current of $8000 \text{ MW} / 500 \text{ kV} = 1.6 \times 10^4 \text{ A}$, or 16 000 A. This is much more feasible. Any practical power line has a significant resistance, which causes an I^2R power loss. 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.

Worked example 10.7A

8000 MW is to be transmitted from the Latrobe Valley to Melbourne along a power line with a total resistance of 1Ω .

- a What would be the total transmission power loss if the voltage at Melbourne was to be:
 - i 250 V?
 - ii 500 kV?
- b What voltage would be needed at the Latrobe Valley end of the line to achieve these voltages?
- c How would the answers to parts a (ii) and b change if the resistance of the power line was halved?

Solution

- a i As in the previous paragraph, the current needed to transmit the power at 250 V would be 32 million amps ($3.2 \times 10^7 \text{ A}$). Thus the I^2R power loss, would be $(3.2 \times 10^7 \text{ A})^2 \times 1 = 1.02 \times 10^{15} \text{ W}$, or about 10^9 MW , vastly more than is to be delivered!
- ii This time the current required is 'only' 16 000 A and so the power loss is $16\,000^2 \times 1 = 2.56 \times 10^8 \text{ W}$ or 256 MW. This sounds like a lot of power, but it is only 3% of the power delivered, a much more feasible situation.
- b To determine the voltage at the supply end we need to calculate the voltage drop along the transmission line. This is given by $\Delta V = IR$. As $R = 1 \Omega$ in each case, the voltage drops are $3.2 \times 10^7 \text{ V}$ and 16 000 V (16 kV) respectively. The first is 32 million volts—totally out of the question of course! In the second case the supply voltage must be $500 \text{ kV} + 16 \text{ kV} = 516 \text{ kV}$.

- C** As the power loss is directly proportional to the resistance, the power loss (at 500 kV) would halve to 128 MW. The voltage drop would also halve and so the supply voltage would need to be 508 kV. 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 $1\ \Omega$ aluminium conductor 100 km long needed to carry the power in this example would be about 6 cm. You need to know that the resistivity, ρ , of aluminium is $2.8 \times 10^{-8}\ \Omega\text{ m}$ and that $R = \rho/LA$.

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 ($\pm 340\text{ V}$) 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 10.30, 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.

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!

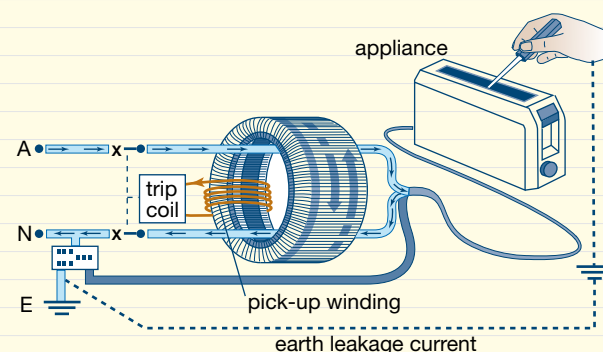


Figure 10.30 Residual current devices prevent fatal shocks by shutting off the current within about 0.03 s.

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 10.31 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 hydro-electricity for rapid response (it is just a matter of turning on the tap) and gas-powered generators (which still take some time to heat up) for the intermediate fluctuations in load. Another way of coping with the fluctuations is to use excess power for 'pumped storage'. The Snowy Mountains Hydroelectric Scheme uses

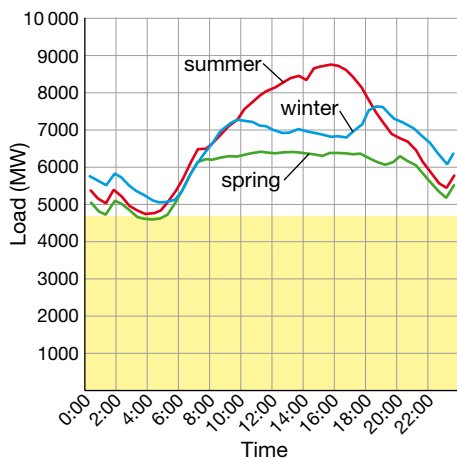


Figure 10.31 Typical load curves for Victoria on a spring, summer and winter day. Each square represents 2000 MW h of energy.

Physics file

The kilowatt hour is simply another energy unit. Energy is the product of power and time:

$$E = Pt$$

1 joule = 1 watt \times 1 second

1 kilowatt hour = 1 kilowatt \times 1 hour

To convert kilowatt hours to joules:

$$\begin{aligned} 1 \text{ kW h} &= 1000 \text{ watts} \times 3600 \text{ seconds} \\ &= 3.6 \times 10^6 \text{ J} \\ &= 3.6 \text{ MJ} \end{aligned}$$

or 1 MW h = 3.6 GJ

around 100–200 MW of excess Victorian coal power each night to pump water back uphill to the high level dams, where it can be re-used to provide power for the morning peak.

Worked example 10.7B

- How much electrical energy was used on the summer day shown in Figure 10.31?
- 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 brown coal is 9 GJ per tonne and each tonne of brown coal burnt releases 0.9 tonne of CO_2 .)

Solution

- The total energy used is represented by the area under the graph. On this graph, each square represents $1000 \text{ MW} \times 2 \text{ h}$, or 2000 MW h.
One megawatt hour = $10^6 \times 3600 = 3.6 \text{ GJ}$. So the energy represented by each square on the graphs is $2000 \times 3.6 = 7200 \text{ GJ}$ or $7.2 \times 10^{12} \text{ J}$.
The number of squares under the summer graph is approximately 80, so the total electrical energy used that day was $80 \times 7200 = 576 \text{ 000 GJ}$, a very large amount of energy!
- At an efficiency of 25%, the total amount of coal energy required is:
 $4 \times 576 \text{ 000} = 2.3 \text{ million GJ}$
Each tonne of coal provides 9 GJ of energy, so the mass of coal required is:
 $2.3 \times 10^6 / 9 = 256 \text{ 000 tonnes}$
(This would fill the MCG to a height of over 2 m!)
The mass of carbon dioxide released from burning this coal would be $0.9 \times 256 \text{ 000} = 233 \text{ 000 tonnes}$. This is equivalent to over 46 million '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 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.

The future

We don't 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 hydroelectricity, 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.

The great challenge for the 21st century is to produce electrical energy without the emission of damaging amounts of the greenhouse gas, carbon dioxide. Unfortunately Victoria uses brown coal, one of the dirtiest forms of coal, in one of the dirtiest coal-fired power stations in the developed world—Hazelwood—and so there is an urgent need to look for better forms of energy and at ways of reducing emissions from coal.

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.



Figure 10.32 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.

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 *inverter*), 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



Figure 10.33 At the heart of an AC-DC converter station are valves comprising power thyristors, which are basically large high voltage transistors.

electromagnetic radiation from the lines. All these factors limit AC transmission lines to a few hundred kilometres.

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 inverter 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 New South Wales, hydro from Tasmania, wind from western Victoria and solar from Melbourne rooftops, the need for large baseload stations is reduced considerably.

Physics in action

Light—an electromagnetic wave

In our course we have frequently heard that light is an electromagnetic wave, but how do we know this? The story of the discovery of the nature of light has been one of the most fascinating aspects of physics since Newton suggested that it was made up of particles while Huygens argued that it was a wave. As we now know, they were both partly right! Young’s interference experiments convincingly demonstrated the wave nature of light, but left unanswered the question as to what exactly was doing the waving. It was Faraday who first suggested that light might in fact have an electromagnetic nature—but he had no way of showing this.

In 1873, James Maxwell published his *Treatise on Electricity and Magnetism* in which he distilled much of the work of Faraday and others into four elegant equations, now known as Maxwell’s equations. The first relates the electric field to the charge that produces it. The second expresses the fact that all magnetic fields are continuous in nature, there being no magnetic ‘charge’, while the third is basically Faraday’s law of electromagnetic induction. Rather than giving an EMF, the third equation relates the changing magnetic field to an electric field—which is what gives the charges their energy (as expressed by the EMF). The fourth relates the magnetic field to the current producing it, but has a very interesting extra term. As well as a term involving the electric current there is a term that involves the rate of change of electric flux. (Like magnetic flux, electric flux is the product of the electric field and the perpendicular area.) This term would not have surprised Faraday, as he suspected, from symmetry, that just as a changing magnetic flux produces an electric field (third equation), so should a changing electric flux produce a magnetic flux.

Maxwell realised that there was an interesting possibility here. Could a situation arise where a changing magnetic flux led to a changing electric flux, which produced a further changing magnetic flux, which ... and so on? He found that indeed there was. His equations showed that oscillating electric and magnetic fluxes could ‘self-propagate’ through space. Furthermore, they would propagate through space at just $3 \times 10^8 \text{ m s}^{-1}$ —the speed

of light! Now was this just a coincidence or was light itself an electromagnetic wave?

There was little doubt in Maxwell’s mind. He had achieved what physicists had been trying to do since Newton; he had found what it was that was waving in light waves—crossed electric and magnetic fields propagating through space at right angles to each other and to the direction of travel.

But what started these oscillating electric and magnetic fields off in the first place? Maxwell’s equations had the answer! Whenever an electric charge *accelerates*, a sort of ‘kink’ arises in the electric field it produces. As well, because the moving charge represents a current, a magnetic field is produced. Sure

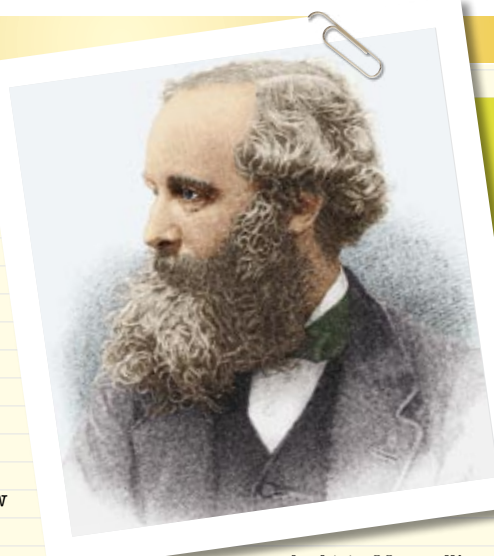


Figure 10.34 James Clerk Maxwell (1831–1879) was born the year Faraday discovered electromagnetic induction and died the year Einstein was born.

$$\oint \vec{E} \cdot d\vec{A} = q/\epsilon_0$$

$$\oint \vec{B} \cdot d\vec{A} = 0$$

$$\oint \vec{E} \cdot d\vec{s} = -\frac{d\Phi_B}{dt}$$

$$\oint \vec{B} \cdot d\vec{s} = \mu_0 \epsilon_0 \frac{d\Phi_E}{dt} + \mu_0 i$$

Figure 10.35 This is a version of Maxwell’s famous equations. We do not need to concern ourselves with the maths! You will see, however, that they are not all that frightening.



Figure 10.36 In 1888, some 20 years after Maxwell predicted them, Heinrich Hertz demonstrated the existence of electromagnetic waves by transmitting them from oscillating currents in a coil to an 'antenna' that produced a tiny spark in a gap.

enough, the changing electric field (the 'kink') is just what is needed to produce the changing magnetic field and the process continues. This was well before Rutherford's atomic model with its accelerating electrons, but it was well established that

matter contained charged particles, and Maxwell suspected that somehow the processes that produced light (heating a wire for example) involved the acceleration of these basic charged particles. This was not unreasonable, as all such processes require the use of energy, and the energy was presumably responsible for the acceleration.

As is often the case in physics, one good discovery opens the way to many more. If atom-sized accelerating charges could produce electromagnetic waves, what about accelerating charges in ordinary sized objects? It was not long before experimenters were trying out the idea. You can probably guess what they found! In 1888, Heinrich Hertz demonstrated that indeed oscillating currents in a coil produced electromagnetic waves that could be detected in another coil. Radio, and the whole basis of much of our modern technological world, was born.

But that wasn't all! Maxwell's equations had predicted the speed of light very accurately. However, they also suggested that it would *always* have this value, regardless of the motion of the source or detector. In other words, it had an *absolute* velocity. However, this conflicted badly with Newtonian physics, which said that all motion should be relative, that there was no possibility of an absolute velocity. Most physicists thought that there must be some mistake in the equations and that eventually this would become clear. There was one young physicist, however, who felt very strongly that Maxwell's equations were so elegant, and so beautifully expressed the nature of electricity and magnetism, that they just had to be true. Perhaps, he suggested, there were other assumptions—about the nature of time and space—that were wrong. His name, of course, was Albert Einstein.

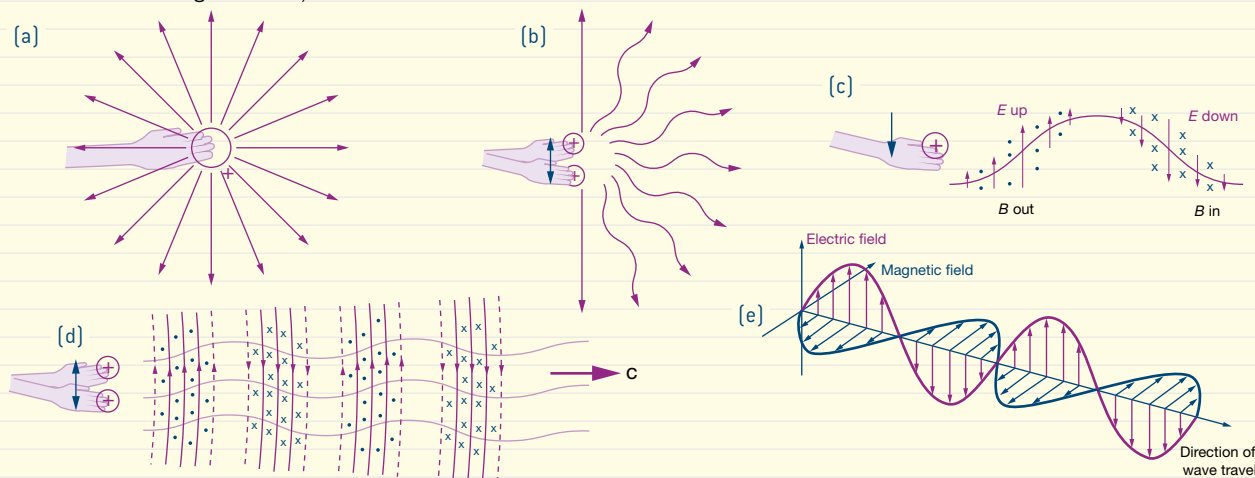


Figure 10.37 Electromagnetic waves are generated when a charge is accelerated, usually by oscillating it.

(a) When a charge is at rest (or moving steadily) it only has a radial (outward) electric field.

(b) When it oscillates, this causes 'wiggles' in the field to move outwards. These wiggles move outwards, just as waves in a rope would if the end were shaken. We are looking only at the waves on the right-hand side.

(c) These waves of field have a tangential component, i.e. at right angles to the radial field. This component is strongest where the wiggle diverges most from radial, i.e. where the charge was moving fastest. This diagram shows the charge at the bottom of its oscillation, having just produced an upward field. A little earlier, as it was moving up, it produced a downward electric field, which has moved further out. A moving charge is a current, and a current produces a magnetic field. So, along with the wiggle in the electric field goes a magnetic field, which will point in and out of the page. As the charge [therefore current] moves down, the field points out of the page. A little earlier it had moved up and produced the inward field shown further out.

(d) A pair of oscillating electric and magnetic fields moves outwards from the vibrating charge. The electric field oscillates between up and down while the magnetic field oscillates between in and out of the page. These two changing electric and magnetic fields are just what Maxwell said would propagate through space at the speed of light. (e) A graphical representation of the situation in (d).



10.7 summary

Using electrical energy

- 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 .
- 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 (or nuclear) plants and the peak demand by hydro or gas turbines.
- Total energy demand is represented by the area under a load–time graph ($E = Pt$).
- The kilowatt hour (kW h) is a useful unit for electric energy. It is equal to 3.6 MJ.



10.7 questions

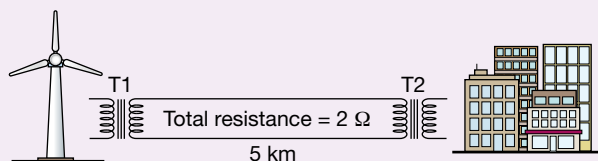
Using electrical energy

All voltages referred to in these questions are RMS.

- In Victoria, electric power is transmitted at very high voltages (up to 500 kV).
 - What is the main reason for this?
 - What factors limit the use of even higher voltages for power transmission?
 - 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:
 - 250 kV?
 - 500 kV?
 - A 100 km transmission line made from aluminium cable has an effective radius of 1.0 cm and a total resistance of $10\ \Omega$. 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:
 - 250 kV
 - 500 kV.
 - A transmission line made from aluminium cable is twice the length (200 km) and twice the radius (2.0 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 5.0 kW of electrical power at 500 V. This power is transmitted to a distant house via twin cables of total resistance $4.0\ \Omega$.
- What is the current in the cables?
 - What is the total power loss in the cables?
 - What is the percentage power loss in the cables?
 - What is the voltage available at the house?
- The supplier is unhappy with this power loss and uses a transformer to step up the transmission voltage to 5.0 kV, and a step-down transformer at the other end. The power output from the generator remains at 5.0 kW and the same $4.0\ \Omega$ cables are used.
 - What current now flows in the cables?
 - Calculate the percentage power loss for this new arrangement.
 - What is the voltage output to the house?
 - If the cost of electricity in Victoria is 15 cents per kilowatt hour, determine the cost of running the following.
 - 1.0 kW heater for 2.0 hours
 - 80 W light globe for 30 minutes
 - 250 W television for 12 hours
 - 6.0 W clock radio for a week
 - Computer printer using 3 W on standby for a year
 - A town 100 km from a power station uses up to 500 MW of power. The power lines between the power station and the town have a total resistance of $2\ \Omega$.
 - If this power was to be transmitted at 250 V, how much current would be required and what would the voltage loss along the power line be? Would this be practical?
 - If the power from the generator is transmitted at 100 kV, what current is required and what would be the voltage drop along the line? What is the voltage at the town?
 - If another power line is added so that the total resistance halves, to $1\ \Omega$, 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 kW generator with an output voltage of 1000 V. The voltage is increased by transformer T1 to 10 000 V for transmission to a town 5 km away through power lines with a total resistance of $2\ \Omega$. Another transformer, T2, at the town reduces the voltage to 250 V. Assume that the transformers are 'ideal'.

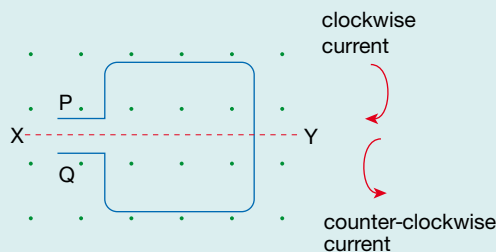


- 9 When the system is running at full power:
 - a what is the current in the power line?
 - b What is the voltage drop along the power line and the voltage at the input to the town transformer?
 - c How much power is lost in the power line? Is this a problem?
- 10 It is suggested that the cost of the first transformer could be avoided if the generator was connected to the power line directly and T2 reduced the voltage at the end of the line to 250 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 T2?
 - 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?

chapter review

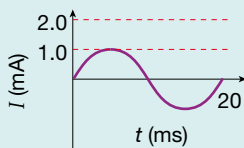
The following information applies to questions 1–3.

A rectangular coil of area 40 cm^2 and resistance 1.0Ω is located in a uniform magnetic field $B = 8.0 \times 10^{-4} \text{ T}$ which is directed out of the page. The plane of the coil is initially perpendicular to the field as depicted in the diagram below.

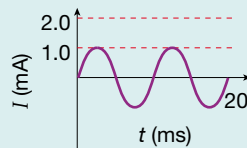


- State the magnitude and direction of the average current induced in the coil when the following changes are made.
 - The magnitude of B is doubled in 1.0 ms .
 - The direction of B is reversed in 2.0 ms .
 - The magnitude of B is halved in 1.0 ms .
- The coil is rotated about the axis XY with constant frequency of $f = 100 \text{ Hz}$.
 - What is the maximum EMF induced in the coil?
 - What is the peak current induced in the coil?
- 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?
 - $f = 50.0 \text{ Hz}$, $B = 8.0 \times 10^{-4} \text{ T}$
 - $f = 200 \text{ Hz}$, $B = 4.0 \times 10^{-4} \text{ T}$
 - $f = 100 \text{ Hz}$, $B = 4.0 \times 10^{-4} \text{ T}$

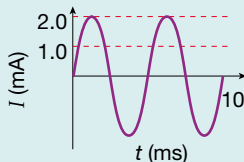
A



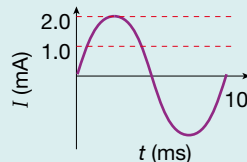
B



C

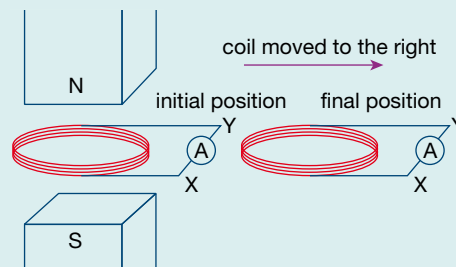


D



The following information applies to questions 4–6.

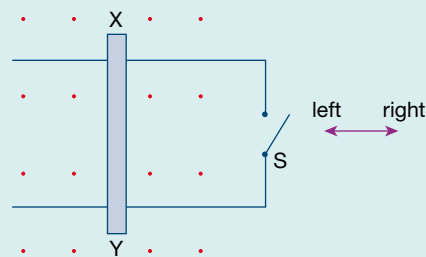
During a physics experiment a student pulls a horizontal coil from between the poles of two magnets in 0.10 s . The initial position of the coil is entirely in the field while the final position is free of the field. The coil has 40 turns, each of radius 4.0 cm , and a total resistance of 2.0Ω . The field strength between the magnets is 20 mT .



- 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?
- 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?
 - The number of turns in the coil is increased.
 - The area of each turn is reduced.
 - The coil is pulled out of the field faster than before.
 - The direction of the magnetic field is reversed before the coil is pulled out.
- 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 7 and 8.

A copper rod, XY , of length 20 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 1.0Ω when it is closed. A uniform magnetic field of strength 10 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 2.0 m s^{-1} .

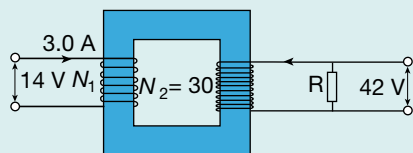


- 7 a What is the magnitude and direction of the average current induced in the rod?
- b What is the magnitude and direction of the magnetic force on the rod due to the induced current?
- 8 The force moving the rod is removed and the rod is stationary. The switch is now opened. What is the direction of the induced current in the rod? Justify your answer.

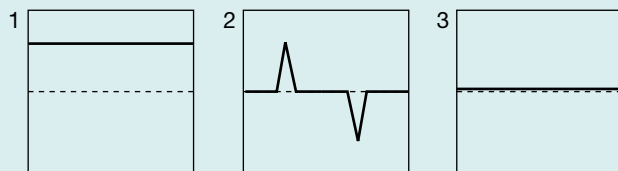
The following information applies to questions 9 and 10.

A ship with a vertical steel mast of length 8.0 m is travelling due west at 4.0 m s^{-1} in a region where the Earth's magnetic field (assumed horizontal) is equal to $5.0 \times 10^{-5} \text{ T}$ north.

- 9 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 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 8.0Ω . What would be the average current induced in the circuit? Justify your answer.
- 10 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 $1.0 \times 10^{-5} \text{ T s}^{-1}$. Calculate the average current induced in the circuit previously described.
- 11 An ideal transformer is operating with an RMS input voltage of 14 V and RMS primary current of 3.0 A. The output voltage is 42 V. There are 30 turns in the secondary winding.

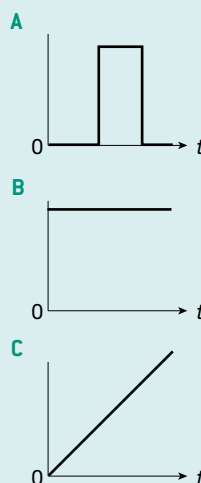


- a What is the RMS output current?
- b How many turns are there in the primary winding?
- c How much power is consumed by the resistor R?
- 12 The following diagrams show the output voltages for a transformer as they appear on the screen of a CRO. The broken line in each display is the time base and represents zero voltage.



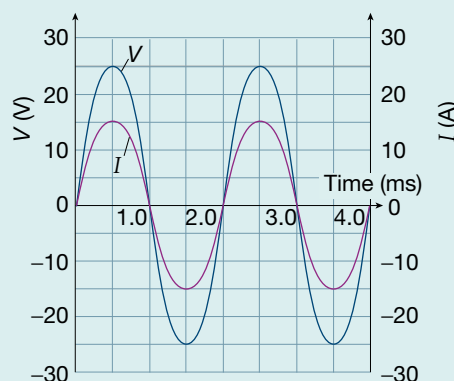
Which of the following input voltages A–C would produce:

- a display 1?
- b display 2?
- c display 3?



The following information applies to questions 13–15.

A physics student uses a CRO to display the current, I , through and the voltage, V , across the terminals of a loudspeaker which has been connected to a signal generator. The graph obtained from the CRO display is shown below.



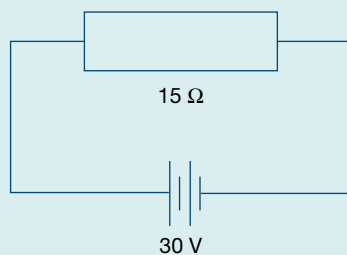
- 13 a What is the frequency of the output signal from the generator?
- b What is the RMS voltage of this signal?
- c What is the peak-to-peak voltage for the signal?
- 14 a Calculate the RMS power output of the signal generator.
- b What is the peak power output of the generator?
- c Calculate the apparent resistance of the speaker.
- 15 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 W. Which set of specifications is consistent with this power output?

Peak–peak voltage (V)

Peak–peak current (A)

A	20	3.0
B	40	6.0
C	40	12.0
D	20	6.0

- 16** Which set of specifications in Question 15 would best describe the AC voltage–current combination that would produce a power consumption equivalent to the DC circuit shown in the following diagram?



- 17** A student builds a simple alternator consisting of a coil containing 500 turns, each of area 10 cm^2 , mounted on an axis that can rotate between the poles of a permanent magnet of strength 80 mT . At a frequency of 50 Hz , it is found that the peak voltage produced is 12.6 V .
- What are the peak-to-peak and RMS voltages?
 - If the frequency is doubled to 100 Hz , how will the peak and RMS voltages change?
 - What frequency of rotation is required for the alternator to produce a voltage of 16 V RMS ?
 - If the magnetic field is reduced to 60 mT , what will be the peak voltage at 50 Hz ?
- 18** A farmer needs to operate a 24 V , 480 W machine that is a long way from the nearest 240 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 2.0Ω . The farmer initially uses the cable to supply the 240 V to the transformer, which is near the machine.

- 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?
- 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 .

- What transformer output voltage will be needed, and what is the turns ratio of the new transformer?
- How much power is being wasted in this arrangement?

The following information applies to questions 19 and 20.

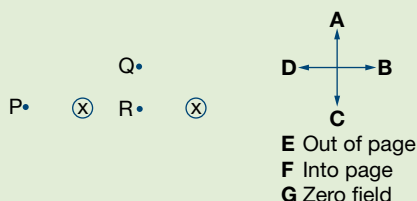
A generator is to be installed in a farm shed to provide 240 V power for the farmhouse. A twin conductor power line with total resistance 8Ω already exists between the shed and house. The maximum power requirement of the house is estimated as 2000 W . The farmer has seen a cheap 240 V DC generator advertised and is tempted to buy it.

- 19** What would you advise the farmer about using the 240 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?
- 20** The farmer is now convinced that he needs an AC generator and that he wants no more than a 5 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?



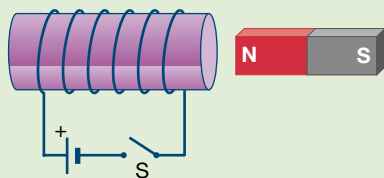
exam-style questions Electric power

- 1 The diagram represents two conductors, both perpendicular to the page and both carrying equal currents into the page. In these questions ignore any contribution from the Earth's magnetic field. The arrows A–D and letters E–G represent options from which you are to choose in answering the following questions.



What is the direction of the magnetic field due to the two currents at:

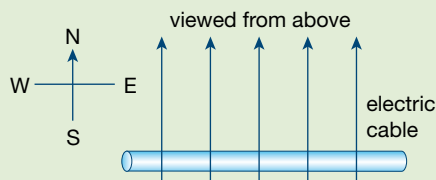
- point P?
 - point Q?
 - point R?
- 2 An electromagnet with a soft iron core is set up as shown in the diagram below. A small bar magnet with its north end towards the electromagnet is placed to the right of it. The switch S is initially open. The following questions refer to the force between the electromagnet and the bar magnet under different conditions.



- Describe the force on the bar magnet while the switch remains open.
- Describe the force on the bar magnet when the switch is closed and a heavy current flows.
- 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 3–7.

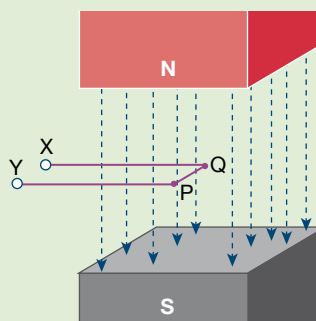
The diagram below shows a horizontal, east–west electric cable, located in a region where the magnetic field of the Earth is horizontal and has a magnitude of 1.0×10^{-5} T. The cable has a mass of 0.05 kg m^{-1} . Assume that $g = 9.8 \text{ N kg}^{-1}$.



- What is the magnitude of the magnetic force on a 1.0 m section of this cable if a 100 A current is flowing through it?
- What is the direction of the current that will produce a force vertically upwards on this cable?
- What magnitude of current would be required to produce zero resultant vertical force on a 1.0 m section of this cable?
- Assume that a 100 A current is flowing through this cable from west to east. What would be the magnitude of the change in magnetic force per metre on this cable if the direction of this current was reversed?
- The cable is no longer horizontal, but makes an angle θ with the direction of the Earth's magnetic field. A 100 A current passing through this cable would produce:
 - the same magnetic force on the cable as when it was horizontal
 - a smaller magnetic force than when it was horizontal
 - a larger magnetic force than when it was horizontal.

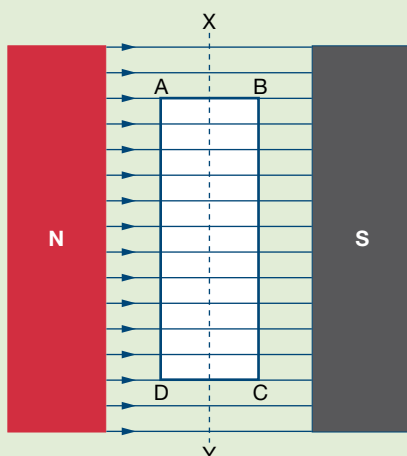
The following information applies to questions 8–11.

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 1.0 T. The side PQ has length 5.0 cm. X is connected to the positive terminal of a battery while Y is connected to the negative terminal. A current of 1.0 A then flows through this loop.



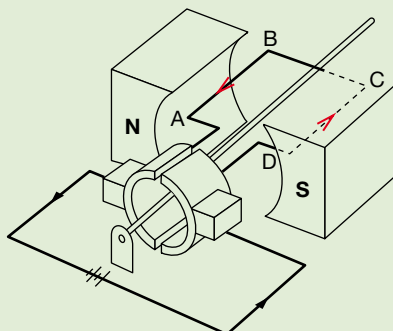
- What is the magnitude of the force on side PQ?
- What is the direction of the force on side PQ?
- What is the magnitude of the force on a 1.0 cm section of side XQ that is located in the magnetic field?
- The direction of the current through the loop is reversed by connecting X to the negative terminal and Y to the positive terminal of the battery. What is the direction of the force on side PQ?

- 12 A rectangular coil containing 100 turns and dimensions $10\text{ cm} \times 5.0\text{ cm}$ is located in a magnetic field $B = 0.25\text{ T}$ as shown. It is free to rotate about the axis XY. The coil carries a constant current $I = 200\text{ mA}$ flowing in the direction ADCB.



- What is the magnitude and direction of the magnetic force on sides:
 - AB?
 - DC?
 - What is the magnitude and direction of the magnetic force on sides:
 - AD?
 - BC?
 - Describe the likely motion of the coil if it is free to rotate.
- 13 A student constructs a simple DC electric motor consisting of N loops of wire wound around a wooden armature, and a permanent horseshoe-shaped magnetic of strength B . He connects the motor to a 9 V battery but is not happy with the speed of rotation of the armature. Which *one or more* of the following modifications will most likely increase the speed of rotation of the armature?
- Increase the number of turns N .
 - Use a 12 V battery instead of a 9 V battery.
 - Replace the wooden armature with one of soft iron.
 - Connect a $100\ \Omega$ resistor in series with the armature windings.

- 14 Consider the electric motor shown.



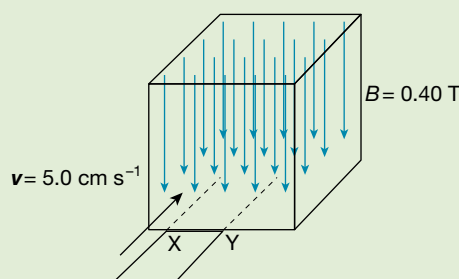
- The direction of the current in the coil is shown (from D anticlockwise to A). What is the direction of the force on sides AB and CD?
- In what position of the coil is the turning effect of the forces greatest?
- At one point in the rotation of the coil the turning effect becomes zero. Explain where this occurs and why the motor actually continues to rotate.

- 15 A rectangular loop of 100 turns is suspended in a magnetic field $B = 0.50\text{ T}$. The plane of the loop is parallel to the direction of the field. The dimensions of the loop are 20 cm perpendicular to the field lines and 10 cm parallel to them.

- It is found that there is a force of 40 N on each of the sides perpendicular to the field. What is the current in each turn of the loop?
- This loop is then replaced by a square loop of 10 cm each side, with twice the current and half the number of turns. What is the force on each of the perpendicular sides now?
- The original rectangular loop with the original current is returned but a new magnet is found which provides a field strength of 0.80 T . What is the force on the 20 cm side now?

The following information applies to questions 16–22.

A square conducting loop with sides 20 cm and resistance $0.50\ \Omega$ is moving with a constant horizontal velocity of 5.0 cm s^{-1} towards a region of uniform magnetic field of strength 0.40 T directed vertically downwards, as shown in the following diagram. The magnetic field is confined to a cubic region of side 30 cm .



- Describe the direction of the induced current in the side XY of the loop just as it begins to enter the field. Justify your answer.
- Calculate the average EMF induced in the loop when it is halfway into the field.
- How much electrical power is consumed in the loop when it is halfway into the field?
- What is the source of this power?
- What is the magnitude of the magnetic force experienced by side XY when it is halfway through the field?

21 What is the average EMF induced in the loop 5 s after it started to enter the cube? Justify your answer.

22 What is the direction of the induced current in the side XY just as it begins to emerge from the field? Justify your answer.

23 A rectangular conducting loop of dimensions $100\text{ mm} \times 50\text{ mm}$ and resistance $R = 2.0\ \Omega$, is located with its plane perpendicular to a uniform magnetic field of strength $B = 1.0\text{ mT}$.

- Calculate the magnitude of the magnetic flux Φ_B threading the loop.
- The loop is rotated through an angle of 90° about an axis, so that its plane is now parallel to B . Determine the magnetic flux Φ_B threading the loop in the new position.
- The time interval for the rotation $\Delta t = 2.0\text{ ms}$. Determine the average EMF induced in the loop.
- Determine the value of the average current induced in the loop during the rotation.
- Will the current keep flowing once the rotation is complete and the loop is stationary? Explain your answer.

24 A $5.0\ \Omega$ coil, of 100 turns and radius 3.0 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 $595\ \Omega$. It is then moved out of the field of the magnet and it is found that an average current of $50\ \mu\text{A}$ flows for 2 s .

- 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?
- What is the strength of the magnetic field?

25 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 field of strength $B = 50\text{ mT}$ and rotated at a frequency $f = 100\text{ Hz}$. It is found that during the rotation, the peak voltage produced is 0.9 V .

- 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.
- What is the RMS voltage generated?
- The student now rotates the coil with a frequency $f = 200\text{ Hz}$. How would your answers to parts a and b be affected?

The following information applies to questions 26–30.

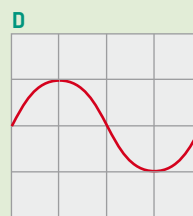
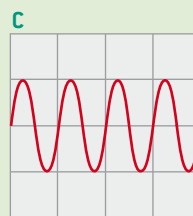
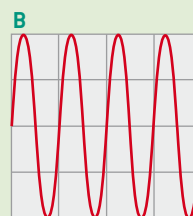
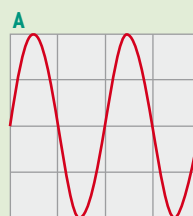
The following diagram represents the alternating voltage output of a generator whose rotor is rotating with a frequency of 50 Hz . The magnetic field strength produced by the magnetising current is 0.50 T . The total number of turns in the armature coils of the stator is $N = 200$, each of area $A = 100\text{ cm}^2$.



26 What is the peak voltage output of the generator?

27 What is the RMS voltage output of the generator?

The following diagrams A–D and table apply to questions 28–30.



	$f\text{ (Hz)}$	$B\text{ (T)}$	N	$A\text{ (cm}^2\text{)}$
A	50	0.50	200	100
B	100	0.50	200	100
C	100	1.00	50	100
D	50	0.50	400	100

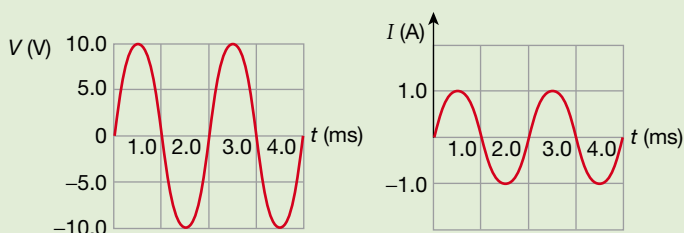
28 Which of the diagrams A–D best describes the display on the CRO when the generator is operating at a frequency of 100 Hz ?

29 Which of the specifications in the table could produce a CRO display described by diagram A?

30 Which of the specifications in the table could produce a CRO display illustrated by diagram C?

The following information applies to questions 31–34.

The following diagram shows the voltage–time graph and corresponding current–time graph for an alternator that was built by a physics student as part of a research project.



- 31 What is the frequency of the voltage produced by the alternator?
- 32 What is the peak-to-peak output voltage of this alternator?
- 33 What is the RMS output voltage of the alternator?
- 34 Calculate the RMS output power of the alternator.
- 35 An ideal transformer is operating with peak input voltage of 600 V and an RMS primary current of 2.0 A. The peak output voltage is 3000 V. There are 1000 turns in the secondary winding.
- What is the RMS output current?
 - What is the output peak-to-peak voltage?
 - How many turns are there in the primary winding?
 - Determine the RMS power consumed in the secondary circuit.
 - Calculate the peak power consumed in the secondary circuit.
- 36 Which of the following is the best description of how a transformer transfers electrical energy from the primary windings to the secondary windings?
- The current through the primary windings produces a constant electric field in the secondary windings.
 - The current through the primary windings produces a steady magnetic field in the secondary windings.
 - The current through the primary windings produces a changing magnetic field in the secondary windings.
- 37 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:
- The primary and secondary coils are in series and so no current can flow in either if the secondary coil is open.
 - There can be no magnetic flux generated in the transformer if the secondary coil has no current in it.

- The magnetic flux generated by the current in the primary produces an EMF that opposes the applied voltage.
- The magnetic flux generated by the secondary coil almost balances out that due to the primary coil.

The following information applies to questions 38–41.

A farmer has installed a wind generator on a nearby hill, along with a power line consisting of two cables with a combined total resistance of $2.0\ \Omega$. The output of the generator is given as 250 V AC (RMS) with a maximum power of 4000 W. She connects up the system and finds that the voltage at the house is indeed 250 V. However, when she turns on various appliances so that the generator is running at its maximum power output of 4000 W, she finds that the voltage supplied at the house is rather low.

- 38 Explain why the voltage dropped when she turned on the appliances in the house. Calculate the voltage at the house.
- She then decides to install transformers at either end of the same power line so that the voltage transmitted from the generator end of the line in this system becomes 5.0 kV.
- 39 Describe the essential features of the types of transformers that are needed at either end of the power line.
- 40 When operating at full load from the generator (4000 W), and assuming ideal transformers, what is:
- the current in the power line?
 - the voltage drop along the power line?
 - the power loss in the power line?
 - the voltage and power delivered to the house?
- 41 How did the power loss in the two systems compare? Explain why it was that the system operated with much lower power loss when the voltage was transmitted at the higher voltage.

Unit

area of study 2

Unit

Interactions of light and matter

outcome

On completion of this area of study, you should be able to use wave and photon models to analyse, interpret and explain interactions of light and matter and the quantised energy levels of atoms.

The nature of light

How incredible it would be if it were possible to put the giants of physics from throughout history together in one room for an hour. Copernicus, Galileo, Newton, Joule, Curie, Heisenberg, Einstein and more. The most likely outcome is that the hour would be spent in heated argument. Let's imagine that just one question is posed to them: *What is light?* None would give the same answer as another. Each would have understandings linked to their era. But what would a great mind like Newton have been able to discover had he had available the technology of today?

All of the great physicists had something of the pioneer and the radical in them. Each came across observational data that either defied established theory or had 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 the nature of light, including:

- diffraction and the effects of the ratio λ/w on the diffraction pattern
- Young's double-slit experiment as evidence for the wave nature of light
- path difference and the constructive and destructive interference of waves
- the qualitative effect of wavelength, distance of screen and slit separation on interference patterns
- the photoelectric effect as evidence for the particle-like nature of light and counterevidence for the wave model of light.

11.1 Review of light and waves

Here we will review some of the key ideas about light that form the background to this area of study.

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 11.1). 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.

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 (Figure 11.2).

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 the particles of the medium experiencing a single disturbance as a pulse passes by, the particles will undergo continual vibration about a *mean position*. Many periodic waves in nature are *sinusoidal* waves, represented by the familiar sine and cosine graphs. Periodic waves are characterised by several features whose definitions follow.

Describing waves

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 , is the time taken for one cycle to be completed—it is measured in seconds (s). If a vibrating source is completing 10 cycles per second and producing a progressive wave across the surface of water, the wave would have a frequency of 10 Hz and a period of 0.1 s.

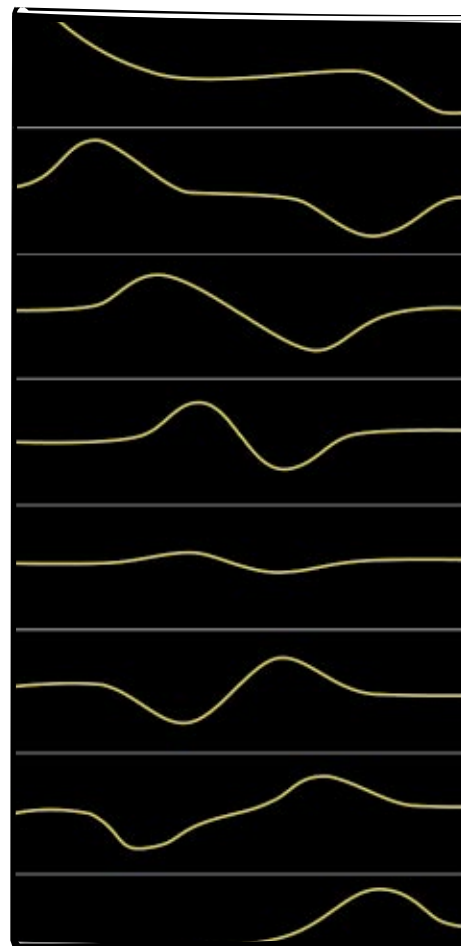


Figure 11.1 A single wave pulse carries energy along the length of the spring with no net transfer of matter.



Figure 11.2 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.

Physics file

Sound waves travelling in air are actually longitudinal mechanical waves where the air molecules make up the medium. Picture the stretched slinky that was discussed. Had the pulse involved an initial disturbance of the spring along its axis, a **longitudinal wave** would have been produced, as the resulting vibration would be parallel to the direction of travel of the wave.

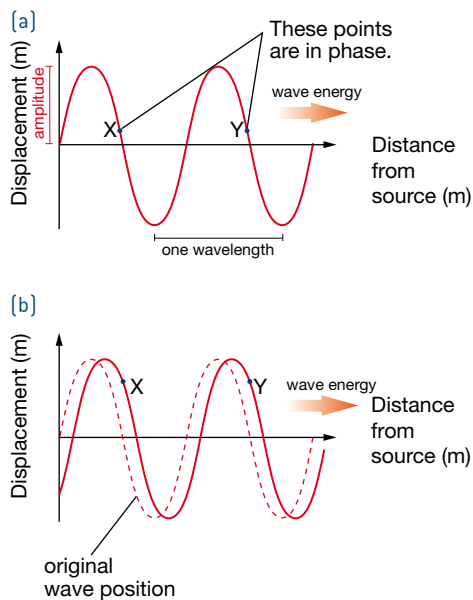


Figure 11.3 As this transverse wave moves away from its source each point in the medium experiences a displacement perpendicular to the direction of travel of the wave. The points X and Y have moved upwards during the period shown.



As their definitions imply, the relationship between the frequency and period of a wave must be:

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

where f = the frequency (Hz)
 T = the period (s)

Consider a medium in which a periodic wave has been present for some time, such as the surface of water. At a particular moment in time, there will be points in the medium which will be experiencing displacements and velocities identical to one another. These points are described as being **in phase** with each other (Figure 11.3). The property of **wavelength** (λ) is defined as the minimum distance between points in a wave which are **in phase**. Alternatively, wavelength can be thought of as the *distance that a wave travels during one period*. Waves can be represented graphically. The wave in Figure 11.3 is moving to the right, i.e. further from the source. If scales are placed on the axes, the amplitude and wavelength can be read directly from this type of graph. Note that the amount of energy that the wave carries with it is a function of the amplitude.

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, the velocity of any wave is given by the expression $v = \lambda/T$ and since $T = 1/f$ we can write:



THE WAVE EQUATION is given by:

$$v = f\lambda$$

where v = the velocity (m s^{-1})
 f = the frequency (Hz)
 λ = 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*. A medium can have a set velocity for each frequency of light. Therefore, this wave 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 11.1A

Waves are created on the surface of water in a tank by a device that dips up and down 20 times per second. The velocity of the resultant wave is 30 cm s^{-1} .

- State the frequency, period and wavelength of the observed wave.
- What happens to the value of the velocity and wavelength if the frequency of the source is doubled?

Solution

- Since frequency is defined as the number of waves or cycles that pass a given point per second, $f = 20 \text{ Hz}$.

Given that $f = \frac{1}{T}$, then:

$$T = \frac{1}{f} = \frac{1}{20} = 0.05 \text{ s}$$

The velocity must be converted to metres per second: $v = 30 \text{ cm s}^{-1} = 0.30 \text{ m s}^{-1}$

Since $v = f\lambda$, we can rearrange to give:

$$\lambda = \frac{v}{f} = \frac{0.30}{20} = 0.015 \text{ m} = 1.5 \text{ cm}$$

- The velocity is determined by the properties of the medium and so it is unchanged, $v = 30 \text{ cm s}^{-1}$. At a set velocity, doubling the frequency will result in wavelengths of half the original value, i.e. 0.75 cm .

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 BC), 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 AD 150. 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.

Reflection and refraction of particles

Recall from your previous studies of light that whenever reflection occurs, the angle of incidence always equals the angle of reflection. In addition, the light reflects in such a way that the incident ray, the normal and the reflected rays all lie in the same plane (Figure 11.4a). This behaviour of light can be accurately modelled using *either* a wave or a particle approach.

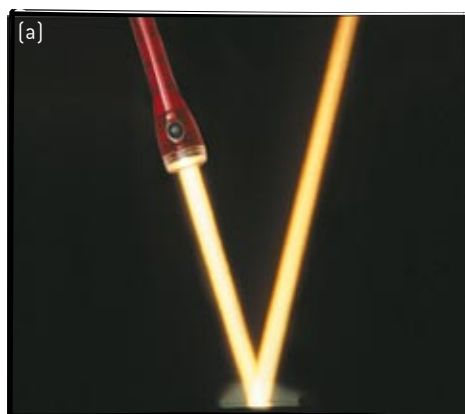
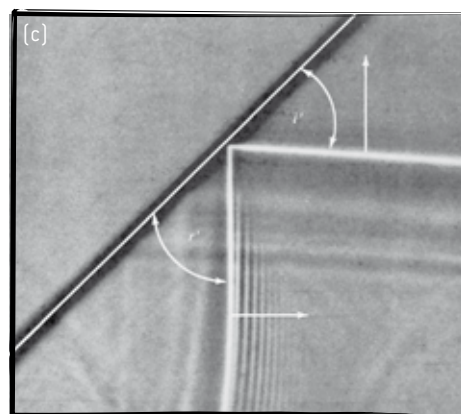
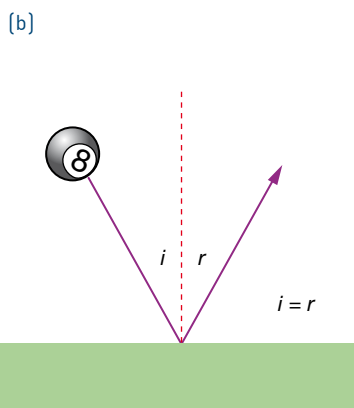


Figure 11.4 (a) The reflection of light can be accurately modelled by either a particle or a wave approach. (b) A billiard ball rolling without spin or sliding will bounce back from the cushion, obeying the law of reflection. (c) Light reflects in the same way as water waves striking a barrier



If light is to be considered to be made up of particles, then these particles can be modelled as behaving just as a billiard ball does when it bounces against the cushion of a billiard table (Figure 11.4b). Light rays were considered to be made up of tiny particles moving at a high speed. Newton, in his influential work *Optiks*, extensively developed a corpuscular model for light (see Physics in action 'Newton's corpuscles').

Alternatively, if light is thought of as a series of wavefronts, reflecting light behaves just as water waves do when striking a barrier (Figure 11.4c). The laws of reflection are obeyed and the reflection of light is modelled just as effectively as with the particle approach. (Note that an important theory about the propagation of waves in a medium was developed by Christian Huygens in 1678. This is discussed later.)

Physics in action

Newton's corpuscles

While considering the problem of chromatic aberration (a colour distortion of images formed by single lenses) in the telescopes of the day, Newton devised a theory for the nature of light that also attempted to explain its colour. By passing a beam of white light through a prism, Newton showed that white light is a mixture of colours. At the time, this was a significant breakthrough, since most scientists of the day thought that colours were mixtures and that white light was pure. As the white light passed through the prism, Newton interpreted the coloured rays of light as streams of tiny particles or 'corpuscles' travelling at high speed that were being sorted by the prism.



Figure 11.5 One of Isaac Newton's many significant contributions to science was the work he did on the phenomenon of dispersion, basing it on his corpuscular model for light.

Newton's corpuscular idea, first published in 1671, drew heavily on the very significant work he had completed in mechanics. If light were a stream of tiny particles, like bullets from a machine gun, then the model could explain why light would travel in *straight lines* and cast shadows. Newton was also able to use this particle idea to show how light might *reflect* and *refract*, obeying the experimental laws that describe each phenomenon.

Newton further developed his model of light so that it could explain other phenomena, such as diffraction (discussed

in section 11.2). To explain diffraction, he argued that a 'corpuscle' of light could change direction as it passed through a small slit or around an obstacle as a result of the forces acting between the corpuscles and the material surrounding the hole.

Although Newton's model could explain many phenomena, there were some it could not. For example, what property of the corpuscles gave the corpuscles their colour? Why could two beams of light cross without the corpuscles colliding and scattering in all directions?

Physics in action

Huygens's wavelets

At about the same time that Newton developed his corpuscular theory, the Dutch mathematician Christian Huygens published his ideas on the nature and propagation of light (1678). His idea was that light acted like a wave. In his model he suggested that each point along a wavefront of light can be considered to be a point source for small, secondary wavelets. Each wavelet is spherical, and the wavelets radiate from their point source in the general direction of the wave propagation (i.e. the light beam). Huygens said that the *envelope* or common tangent of the wavelets becomes the new wavefront.

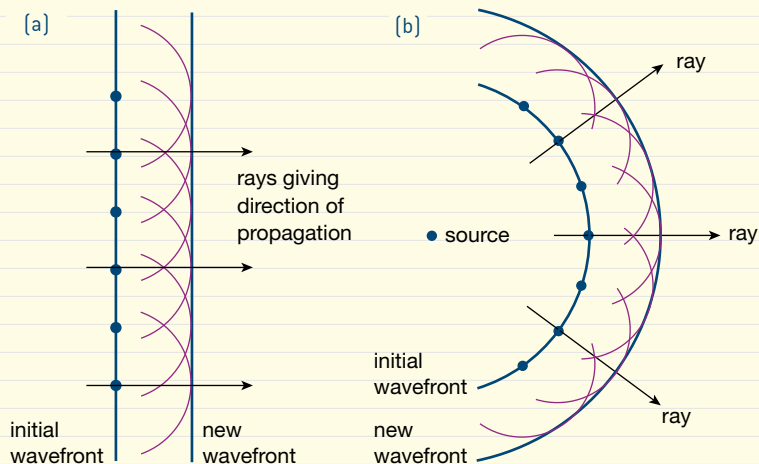


Figure 11.6 Huygens's principle models the propagation of waves. (a) Each point along the plane wavefront is a source of secondary spherical wavelets. The envelope of the wavelets causes the formation of the new wavefront which is also plane. (b) The same treatment illustrates how a circular wave is reproduced.

Huygens's model for light easily accounts for the propagation of both plane and circular wavefronts (Figure 11.6). The sequence of diagrams in Figure 11.7 show how Huygens's principle can be used to explain why the angle of reflection of a light wave is the same as the incident angle. As points along the wavefront meet the mirror they become sources of wavelets emerging upward. The envelope of the new wavelets form the reflecting wavefront.

Refraction

From your earlier studies of light, you should recall that light travels in a straight line when it is travelling in a uniform medium, but if the medium is not uniform, the path of light can be bent. For example, the bending of light by layers of air of different density produces the mirages seen in the desert. The bending of the path of light due to a change in speed as it enters a medium of different optical density is called *refraction*.

Newton's particle approach to modelling refraction was only partially successful. Keep in mind that at this time the speed of light in different media had not been determined. If considering light as a particle, as Newton did, it can be argued that if a ball rolls along a flat plane with a constant speed v_1 , and then rolls down a steep slope to another flat plane with a higher speed v_2 , the direction in which the ball travels will change. The path of the ball will move closer to the normal to the boundary.

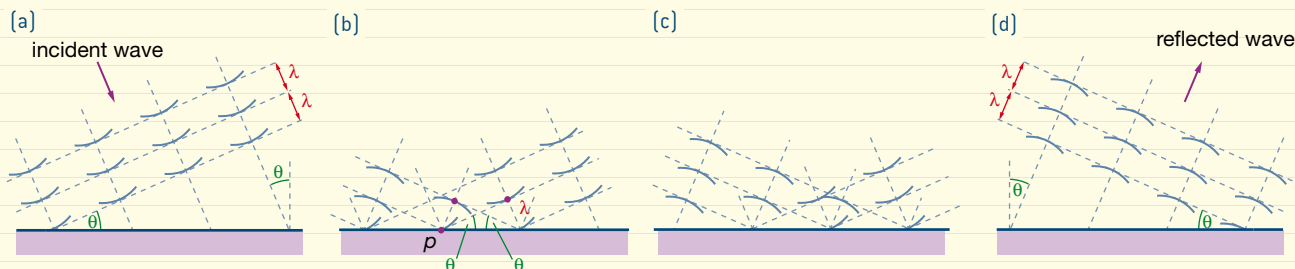


Figure 11.7 Huygens's construction of the reflection of light as a plane wave.

Conversely, if the incline between the flat regions is uphill, then v_2 will be lower than v_1 , and on the upper region the ball will move along a straight-line path further away from the normal. In both cases, Snell's law of refraction would be obeyed by moving particles. The path would be bent through an angle of the correct *size*. (Dutch mathematician Willebrod Snell van Royen had devised the law in 1621.) However, later work on measuring the actual speed of light in different media showed that Newton's *direction* of bending was incorrect. Light was found to bend *towards* the normal when it *slowed down*.

As with reflection, Huygens's ideas could also explain the refraction of light by tracking the new wavefront's position by relying on the idea of secondary wavelets. Figure 11.8 shows a wavefront incident on a boundary to a medium in which the wave travels more slowly. As the points along the wavefront that are meeting the boundary become sources of Huygens's secondary wavelets, these wavelets now travel in a slower medium and so cover less distance in a given time. As a result

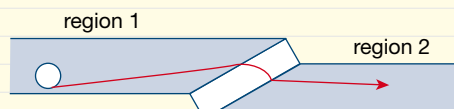


Figure 11.8 If Newton's model was correct, light should speed up as it enters glass from air, just as the rolling ball speeds up as it reaches a downward slope, and produce refraction towards the normal. The first measurements of the speed of light in air and water from 1849 onwards showed that as light speeds up, the bending is in the opposite direction to Newton's prediction. This showed a major flaw in the particle model for the refraction of light.

of this, the joining of the envelope of new wavelets shows that the overall wavefront is heading in a slightly altered direction (Figure 11.9).

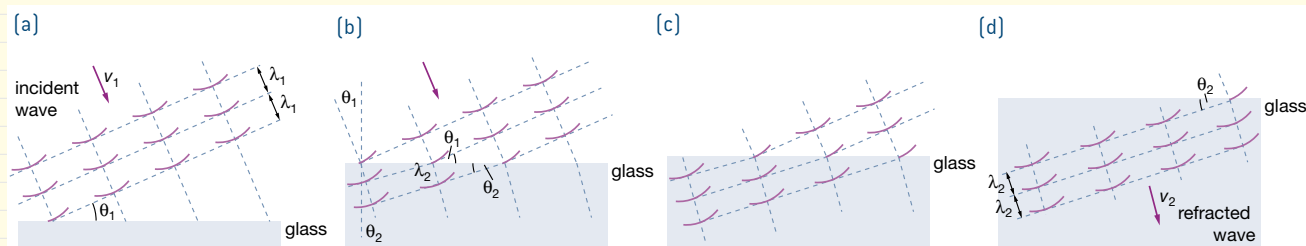


Figure 11.9 Huygens's principle can be used to accurately model the refraction of light and it also explains the subsequent change in wavelength of the light.

Competing ideas

For about a century, the ideas expounded by Newton and Huygens (see previous Physics in actions) vied with each other for acceptance in scientific circles. Newton's corpuscular theory was the more popular, and most of the progress in the 18th century was made with his theory. Newton's corpuscles were thought to be the more logical construction, and Newton's prestige added weight to his arguments. After all, he had been proved right so many times before—why should he be wrong now?

It seemed that the only way to separate these competing theories was to actually determine the speed of light in both air and another transparent medium, such as glass. If light was slower in air, then the particle theory would be considered correct. If the light was faster in air, Huygens's wave model would be vindicated. In fact, it was not until 1850 that these measurements could be made. By then, however, the wave nature of light was well established, and Newton's corpuscles were all but forgotten.

There is not room to fully discuss all of the behaviours of light that have been modelled at some stage in history. Some of them are listed in Table 11.1. However, keep in mind that the *properties of light* exist independently of the model that we choose to represent light. Should we allow ourselves to be constrained in our understanding of this entity by our ability to construct adequate models?



PRACTICAL ACTIVITY 36

Newton's particle model of light

Table 11.1 A summary of light behaviour

Behaviour of light satisfactorily modelled by particles	Behaviour of light satisfactorily modelled by waves
Linear propagation	Linear propagation
Intensity reduces with the square of the distance from the source	Intensity reduces with the square of the distance from the source
Reflection	Reflection
	Refraction
Modelling of light paths around mirrors and lenses with geometric optics	Modelling of light paths around mirrors and lenses with wave optics
	Light beams crossing paths undisturbed
	Partial reflection and transmission of light at a boundary
	Diffraction (see section 11.2)



11.1 summary

Review of light and waves

- 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 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 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:

$$v = f\lambda$$

where v is the velocity (m s^{-1}), f is the frequency (Hz) and λ is the wavelength (m).

- In the 17th century, two models for the nature of light competed with each other for acceptance. Newton's 'corpuscular' model considered light to be composed of small particles, while Christian Huygens promoted a wave model based on wavelets.



11.1 questions

Review of light and waves

- What do mechanical and electromagnetic waves have in common?
- Suggest why mechanical waves generally travel faster in solids than in gases.
- A guitar string is plucked and it undergoes 40 vibrations in 0.25 s. What is the frequency and period of the sound wave produced?
- A student sending transverse waves along a slinky spring completes three full cycles in 1 s. The wave train produced has a total length of 1.80 m. Determine the speed of the wave in the spring.
- What is the wavelength of a sound wave with a velocity of 345 m s^{-1} and a frequency of 1200 Hz?
- 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.
- Which one of the following is correct? The corpuscular model of light, as suggested by Newton, predicts that light should travel:
 - faster in air than in a vacuum
 - faster in a vacuum than in air
 - at the same speed in both media.
- Which one or more of the following is explained successfully only by a wave model for light?
 - Straight-line propagation of light
 - Snell's law
 - The simultaneous reflection and refraction of light at an interface
 - Diffraction and interference
 - The existence of light pressure
- Sometimes when you look into a shop window, you can simultaneously see both the items on display and an image of yourself in the glass. Why was this phenomenon more easily explained by a wave model than by a particle model for light?

11.2

The wave model established



PRACTICAL ACTIVITY 37

Diffraction of light

Diffraction

If you have ever watched carefully as water waves lap gently against a partial barrier, you may have noticed their ability to bend around and enter the region behind the barrier. An example of this is shown in Figure 11.10. This phenomenon is known as **diffraction**. As it is an observed property of both water and sound waves, we will see later that it contributed to the acceptance of the *wave* nature of light.



Figure 11.10 Water waves will bend around an obstacle. Sound waves diffract as well, allowing us to hear around corners.

It is an observed property of waves that obstacles placed in their path will not produce sharply defined 'shadow' regions; rather the waves can bend around and enter the area behind the barrier. This also applies when waves pass through an aperture, which can be thought of as a pair of barriers, as shown in the ripple tank of Figure 11.11. Figure 11.11a shows that a large region that is behind the barrier has been filled with waves, whereas when the wider slit is used (Figure 11.11b) the waves have not bent around as far behind the edge of each barrier. In the diffraction of waves it is observed that

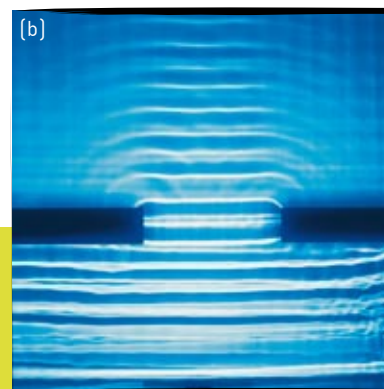
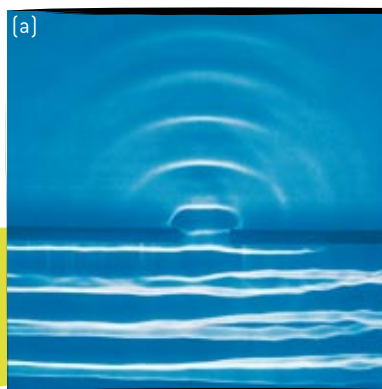


Figure 11.11 The diffraction of water waves in a ripple tank. (a) Significant diffraction occurs when the wavelength approximates the slit width, i.e. $\lambda \approx w$. (b) As the gap increases, diffraction becomes less obvious, since $\lambda \ll w$.

if the wavelength is much smaller than the diameter of the gap or obstacle, the degree of bending is less. Wavelengths comparable to or larger than the diameter of the obstacle or gap will produce significant bending. This can be expressed as the ratio $\lambda/w \geq 1$, which is discussed later in the chapter.

Interference of light

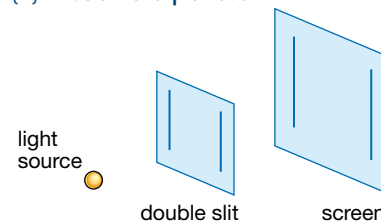
In 1801, an English physician and physicist, Thomas Young, conducted an experiment which showed that light did indeed have a wave nature. In his experiment, Young projected a narrow beam of light through a pair of closely spaced holes in a darkened room. The light from the holes was viewed on a screen some distance away. If the light from the two holes produced a very bright region on the screen where these beams overlapped, Newton's hypothesis—that light was particulate in nature—would be supported. What Young saw in the region where the two beams of light fell, however, was a series of alternating bright and dark fringes (see Figure 11.12).

Young knew about the **interference** of water and sound waves, so he could only conclude that light too must have a wave nature. Young suggested that at each bright band, the light coming from each source was added, and that at each dark band it was being cancelled.

Young's double-slit experiment

To see how this is possible, consider this modern version of Young's experiment. **Monochromatic light** (i.e. light consisting of only one wavelength) is directed at two narrow, parallel slits, S_1 and S_2 , which are placed very close together. A screen is placed at some distance behind the slits. As the light passes through each slit, it diffracts, and two sets of circular wavefronts emerge and spread into the region beyond the slits. Because the

(a) Particle-model prediction



(b) Wave-model prediction

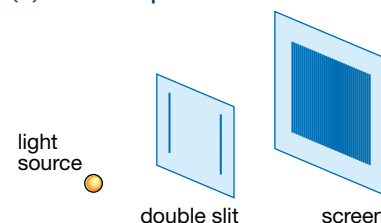


Figure 11.12 (a) In a reproduction of Young's experimental set-up a particle model for light would incorrectly predict that two bright bands would appear on the screen. (b) However, the observed fringe pattern supports a wave model for light.

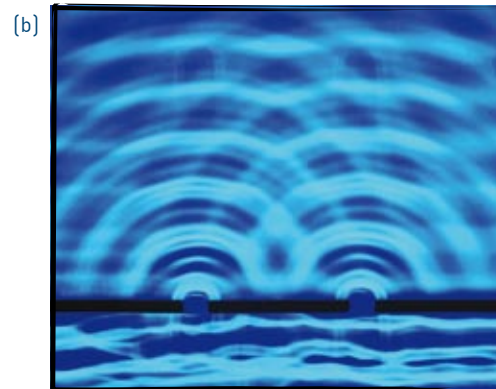
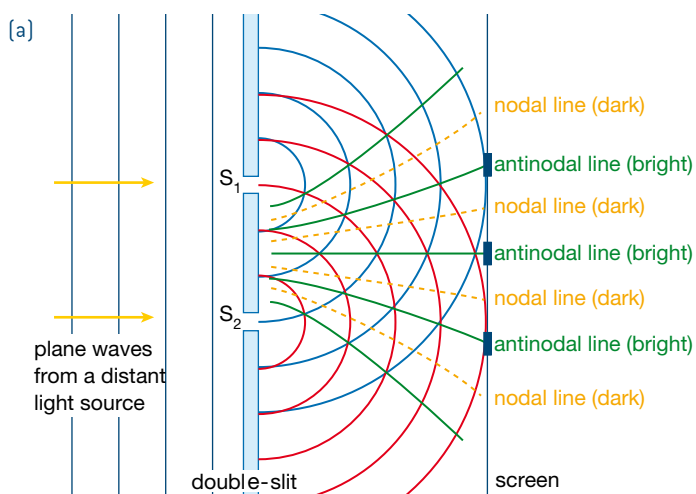



Figure 11.13 (a) Light from a monochromatic and distant source is incident on a pair of slits. Plane waves enter each slit at the same time, resulting in two sets of circular waves that overlap and spread into the region beyond the slits. The circular waves then interfere to produce a series of nodal (dark) lines and antinodal (bright) lines. (b) This interference pattern of alternating nodal and antinodal lines can also be seen in water waves when a pair of dippers oscillate in phase.



incident light consists of plane waves which come from the same source, the light emerging from each slit will be **in phase**. This means that if there is a crest emerging from S_1 , there will also be a crest emerging from S_2 .

At every point on the screen, a wave train from one slit will meet a wave train from the other slit. The resulting intensity at each point is determined by the principle of **superposition**, in which constructive and destructive interference of waves produce nodes and antinodes (see below).

Path difference and interference effects

At a particular point on the screen, P, each wave train will have travelled a different distance, i.e. S_1P and S_2P . The difference in the distance travelled by each wave train to a point P on the screen is called the path difference for the waves (pd).



The **PATH DIFFERENCE** to point P from wave source S_1 and S_2 is given by:

$$pd = |S_1P - S_2P|$$

Path difference can be measured in metres, but it is far more useful to measure it in wavelengths in order to determine the light intensity on the screen.

At the point on the screen equidistant from each slit, M, each wave train will have travelled through the same distance and so there is no path difference (i.e. $S_1M = S_2M$). The light waves **arrive in phase**. These light waves reinforce to produce an **antinode**. A fringe of bright light is seen. This phenomenon is called *constructive interference* or reinforcement.

Reinforcement will also occur whenever the path difference between the two wave trains differs by an integral, or whole, number of wavelengths, i.e. $pd = \lambda, 2\lambda, 3\lambda, \dots$ etc.



CONSTRUCTIVE INTERFERENCE of coherent waves occurs when the path difference

$$pd = n\lambda, n = 0, 1, 2, 3, \dots$$

In Figure 11.14, the path difference at R is λ .

In looking along the screen away from the central bright fringe, there will be a point at which the path difference is $\lambda/2$, i.e. N in Figure 11.14. The two wave trains that meet at this point are **completely out of phase** and cancel each other to produce a **nodal point**. *Destructive interference* is occurring, and no light is seen. Destructive interference also occurs when the path difference between the waves is $3\lambda/2, 5\lambda/2, 7\lambda/2$ etc.



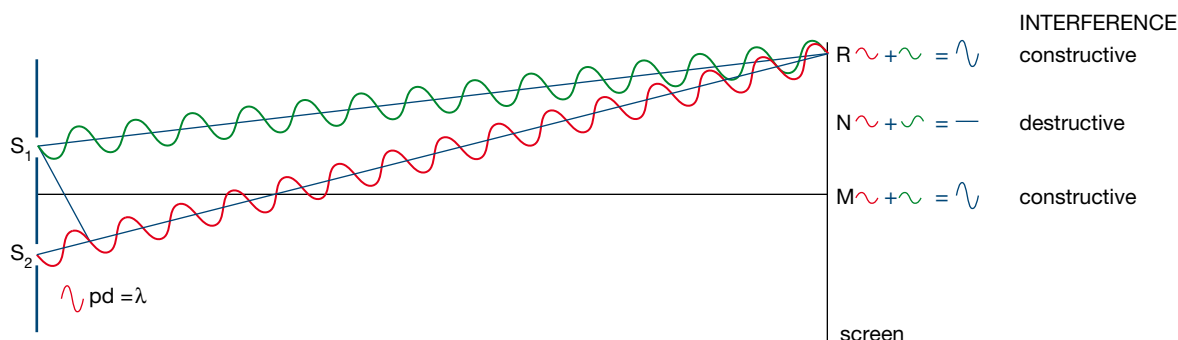
DESTRUCTIVE INTERFERENCE of coherent waves occurs when the path difference equals an odd number of half wavelengths or $pd = (n - \frac{1}{2})\lambda, n = 1, 2, 3, \dots$

The sequence of constructive and destructive interference effects produces an **interference pattern** of regularly spaced vertical bands or fringes on the screen, as shown in Figure 11.15a. This interference pattern can also be represented graphically as shown in Figure 11.15b.



PRACTICAL ACTIVITY 38

Interference of light



Factors affecting interference patterns

The wavelength of light, the separation between the two slits and the distance to the screen can all affect the distance between adjacent bright bands on the screen; that is, the *fringe spacing*. Obviously, if the viewing screen is moved further from the two slits, the fringes will appear further apart from each other. Alternatively, using light of a longer wavelength will result in increased fringe spacing. However, if the screen position and wavelength of light are held constant, surprisingly the fringe spacing can be increased by *reducing* the separation of the slits.



In dual-source interference patterns the **FRINGE SPACING** can be increased by:

- increasing the wavelength
- increasing the distance to the screen
- reducing the slit separation.

Figure 11.14 Waves meeting from each slit at the centre of the screen (M) have travelled through the same distance and, since they are in phase, they reinforce to produce a bright fringe (i.e. constructive interference has occurred). Alongside this 'central maximum' is the first dark fringe, N. The wave train from S_2 will have travelled an extra $\lambda/2$ compared with the wave coming from S_1 , resulting in the two waves being out of phase at all times, i.e. the path difference is $\lambda/2$. Destructive interference is the result. Further out at R, where the path difference is λ , a bright fringe will be seen as the wave trains arrive at this point in phase again. The result is that the interference pattern consists of a pattern of alternating bright and dark fringes.

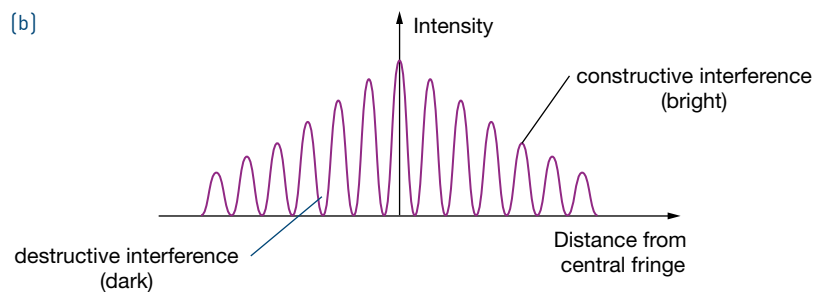


Figure 11.15 (a) The fringes produced by double-slit interference are evenly spaced. The central band is brightest and the adjacent fringes diminish in intensity. The overall width of the pattern is a diffraction effect. (b) The double-slit interference pattern can be considered in terms of an intensity distribution graph. The horizontal axis represents a line drawn across the screen. The centre of the distribution pattern corresponds to the centre of the brightest central fringe.

Different colours (wavelengths) of light produce different fringe-spacing, and Young was the first to determine approximate values for the various wavelengths of visible light. He showed that the physical property distinguishing the different colours of light from each other is actually their wavelength.

Physics in action

Using interference patterns to deduce wavelength

To determine exactly where the fringes of an interference pattern will appear on the screen, a relationship can be determined from the geometry of Young's experiment. Consider wave trains from S_1 and S_2 that travel to a point X on the screen. The wave train from S_2 will travel further than the ray from S_1 , and this extra distance is the path difference. The path difference can be found by constructing a line from S_1 perpendicular to the line bisecting the angle S_1XS_2 . This is the line S_1A on Figure 11.16.

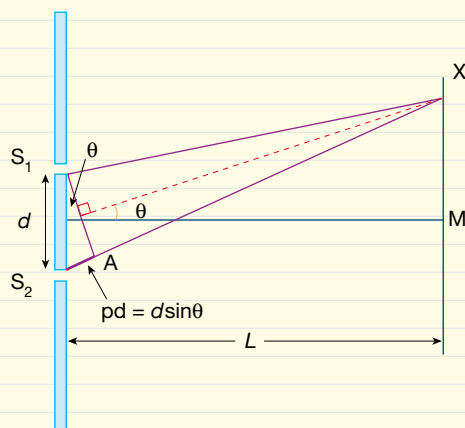


Figure 11.16 The path difference between the wave trains can be determined from a line from S_1 that runs perpendicular through the bisector of S_1X and S_2X . The triangle S_1S_2A is very nearly a right-angled triangle.

Since the distance from the slits to the screen (L) is far greater than the distance (d) between the slits, the fringes will be very close together, and the angle θ will be extremely small. This means that the angle between the wave train S_2X and the perpendicular S_1A will be so close to 90° that it can be treated as a right angle. So $\sin\theta = pd/d$ and thus the path difference is given by $pd = d\sin\theta$.

Now if X is positioned so that it lies at the centre of the first bright fringe after the central fringe, then MX will equal the fringe spacing W , so the path difference will be just λ . This means that $\tan\theta = W/L$, and $\sin\theta = \lambda/d$.

Now, for very small values of θ , $\sin\theta$ approximates closely to $\tan\theta$ (try it with your calculator), so:

$$\tan\theta = \frac{W}{L}, = \frac{\lambda}{d} \text{ and thus } W = \frac{\lambda L}{d}$$

For example, light of an unknown wavelength emitted by a laser is directed through a pair of thin slits separated by $50 \mu\text{m}$. The slits are 2.0 m from a screen on which five bright fringes are seen over a width of 12.5 cm . The wavelength of the laser light can be determined.

We know that $L = 2.0 \text{ m}$, $d = 50 \mu\text{m} = 5.0 \times 10^{-5} \text{ m}$, and $5W = 12.5 \times 10^{-2} \text{ m}$ (so $W = 2.5 \times 10^{-2} \text{ m}$).

Since $W = \frac{\lambda L}{d}$, we have:

$$\begin{aligned} \lambda &= \frac{Wd}{L} \\ &= \frac{2.5 \times 10^{-2} \times 5.0 \times 10^{-5}}{2.0} \\ &= 6.3 \times 10^{-7} \text{ m} \end{aligned}$$

The wavelength of the laser light is therefore 630 nm .

Observing diffraction

The development of our knowledge of diffraction is a very good example of how scientific knowledge often progresses alongside developing technology. Ordinarily, diffraction effects are not observed in everyday circumstances as we do not see diffraction fringes around the shadows of objects unless particular conditions occur. An object or aperture is best illuminated by a **point source** of light, otherwise the diffraction fringes created by each part of a large light source would overlap and uniformly illuminate the area adjacent to the shadow. Similarly, monochromatic light works best. If white light is used, fewer or no fringes may be seen, as the different wavelengths produce fringes of different spacing and these too may overlap. The shadow of a razor blade shown in Figure 11.17a has been produced by using the special conditions described above.

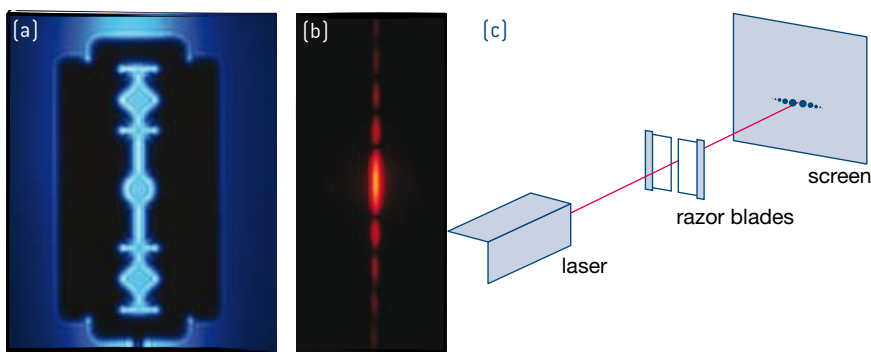


Figure 11.17 (a) Diffraction around a razor blade screen. (b) Diffraction occurs when light passes through a single slit. (c) The extent of the diffraction pattern changes as the distance between the blades changes. If the slit size decreases, the pattern will spread.

Note that the wavelengths of visible light first determined by Young turned out to be far smaller than anyone could imagine. Diffraction effects are significant only if the aperture or obstacle is of a comparable size to the wavelength of the wave passing through. Waves diffract significantly when they interact with objects of a similar size, and even single cells are far larger than the wavelength of light. Since the wavelength of visible light is far smaller than the size of any everyday object, diffraction is not readily seen. As a consequence, beams of light are observed to travel in straight lines and cast sharp shadows. This is why the behaviour of light is, in most circumstances, so adequately modelled via *geometric optics*, i.e. the ray diagrams studied in Unit 2.

To produce an observable example of *single-slit diffraction*, a single aperture can be created by scratching a fine line on a black-painted projector slide, or mounting the edges of two razor blades so that there is only a tiny gap between them (Figure 11.17c). The observed diffraction pattern is once again a series of bright and dark fringes whose existence can be *modelled* by using waves which are said to interfere constructively and destructively, producing the observed pattern on the screen. The extent of the diffraction of the light waves depends on the relative sizes of the wavelength, λ , and the diameter or width of the gap, w (as it does for the water waves discussed earlier). It is useful to consider their relative sizes by referring to the ratio 'wavelength/diameter of slit or obstacle'; that is, λ/w . We say:

Extent of diffraction $\propto \frac{\text{wavelength}}{\text{size of slit}} = \frac{\lambda}{w}$

Since this ratio determines the spacing of the individual fringes, it determines the width of the overall diffraction pattern. The diffraction pattern will have a central fringe that is twice as wide as the other fringes. If the wavelength is held constant and the aperture or gap is made smaller, greater diffraction is seen. If different wavelengths enter the same gap, those with a small wavelength will undergo less diffraction than those with longer wavelengths. For example, the use of an aperture of a given width will result in greater diffraction of red light than blue light just as occurred

Figure 11.18 Since such small apertures are required in order to demonstrate the diffraction of light, red light (a) is diffracted to a greater extent than blue light (b). Red light's longer wavelength results in more widely spaced fringes and a wider overall pattern.

Physics file

Traditionally, seven colours are identified in the visible spectrum: red, orange, yellow, green, blue, indigo and violet. The original seven were identified by Isaac Newton, who felt that there ought to be seven since this was a 'perfect number'!

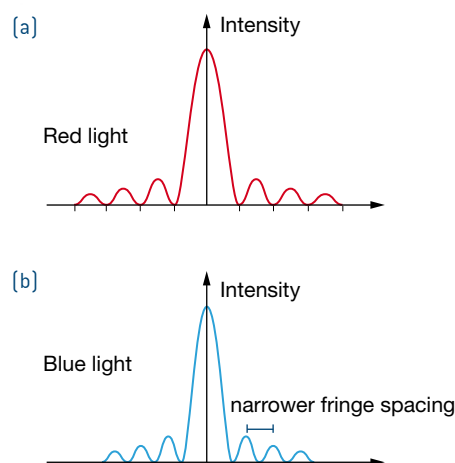
In reality, indigo overlaps the colours violet and blue, and in recent years it has been removed. At birth, the human eye can detect light with a wavelength as low as 280 nm, but this value increases early in life to just below 400 nm.

Table 11.2 Wavelengths of different coloured light

Colour	Wavelength (nm)
Red	750–630
Orange	630–600
Yellow	600–560
Green	560–490
Blue	490–420
Violet	420–400

Physics file

Observe a crude single-slit diffraction pattern by making a 3 cm cut perpendicular to the edge of a small piece of dark (opaque) paper. Hold the paper about 5 cm in front of one eye and look through the slit at a light source such as a computer screen. You should see 'fringes' of light and dark. Pull the paper so that the slit widens and observe the effect.



Physics file

While diffraction effects are often undesirable in optical instruments, we actually want the sound waves put out by our stereo systems to diffract. The spreading out of sound waves is important so that the sound we hear is consistent throughout the room that we are in.

with double-slit diffraction (Figure 11.15). Do not think of diffraction effects as suddenly occurring at a specific wavelength. Although we need $\lambda \approx w$ for a good diffraction pattern, diffraction will gradually increase if the values of λ and w are made to gradually approach one another.

Diffraction effects limit optical instruments

So far, we have discussed that since tiny apertures are required, it is quite unusual to notice the effects of light diffraction in our everyday experiences. However, the diffraction created by the small apertures in microscopes and telescopes results in blurred images. For example, a significant problem is that the light from two tiny objects being observed can be diffracted so much that the two objects appear as one blurred object. We say that the objects are unresolved. Essentially, the same ratio, λ/w , dictates how small an object can be clearly imaged by a particular microscope. A general rule is that optical instruments can only create images of objects of a size similar to the wavelength of the light they use; otherwise, diffraction effects are too significant.

Physics in action

Poisson bright spot

After publishing his results on interference, Young had not won a great deal of support for his ideas in England. In France, however, Augustin Fresnel was developing a mathematical theory for light on the basis of this wave idea. One night, the mathematician Simeon Poisson suggested the following to him. If light were a wave, then parallel waves falling on a solid circular disk would diffract at each point on the edge of the disk, and some waves would be diffracted to the centre of the shadow region. At this point, all the waves would interfere, since they would all arrive in phase (Figure 11.19a).

This idea that there could be light in the middle of a shadow seemed far-fetched, but the experiment was carried out, and a spot of light was seen. The wave model seemed, finally, to be vindicated.

Apart from the bright spot seen in Figure 11.19b, there are a series of bright and dark fringes within the shadow zone. These resemble an interference pattern and are due to the interference of light waves being diffracted around the edges of the disk, creating a two-dimensional diffraction pattern.

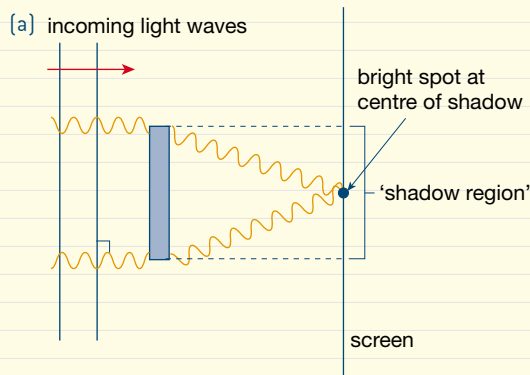
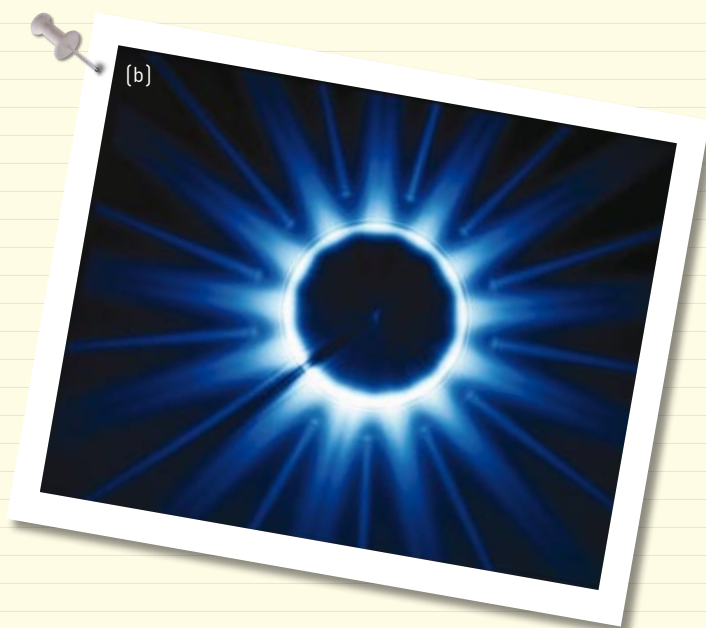


Figure 11.19 (a) Waves of light incident on a solid disk diffract to give a point of light in the centre of the shadow zone. This is convincing evidence for the wave nature of light. (b) Diffraction has produced the Poisson bright spot due to constructive interference.



Diffraction gratings

A **diffraction grating** is a flat piece of glass or plastic which might have as many as 10 000 lines per centimetre. Each line is a scratch that is opaque to any light that is incident on the glass. The gaps between the scratches remain transparent. If polychromatic light (a mixture of colours) is used, each wavelength produces a set of fringes at different angles

to the incident beam. The result is that this light will be separated into its component colours, and a spectrum is seen (Figure 11.20). Diffraction gratings produce such sharp fringes that they are a far more precise device than double slits for determining the wave-length of monochromatic light.

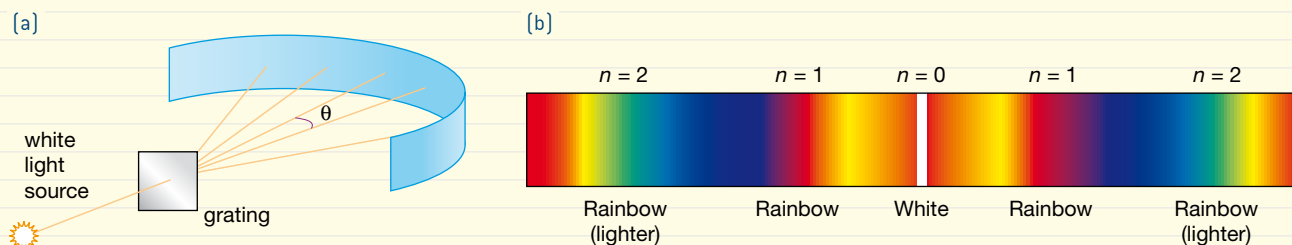


Figure 11.20 (a) The interference pattern from a diffraction grating for white light. The pattern extends through large angles. (b) The maxima are so sharp that the component wavelengths of white light will separate into a spectrum on the screen.

Maxwell's wave theory

In 1864, James Clerk Maxwell produced a complete wave theory for light. In four simple equations, Maxwell showed that electricity and magnetism were both elements in one complete theory—electromagnetism. Maxwell's wave theory demonstrated that all light was **electromagnetic radiation** (EMR) consisting of electric and magnetic fields which oscillate at 90° to each other. Both fields are also perpendicular to the direction of propagation of the light. The frequency of the oscillation of each field is identical, and is equal to the frequency of the light.

Maxwell's theory explained how, once produced, electromagnetic radiation could be self-propagating. This was one of the properties of light which Huygens grappled with some 200 years earlier. In reproducing itself, the electric field in a wave of light creates or induces a changing magnetic field, which in turn recreates the changing electric field an instant later. In this way, a ray of light continues unaltered through space in a straight line until it is either reflected, refracted or absorbed. Some of the light we can see today with powerful telescopes has been travelling through the Universe for more than 10 billion years!

Maxwell's theory was also able to provide a theoretical value for the speed of light, which is the value we accept today: $c = 3 \times 10^8 \text{ m s}^{-1}$. This matched the experimental value determined by Fizeau in 1849.

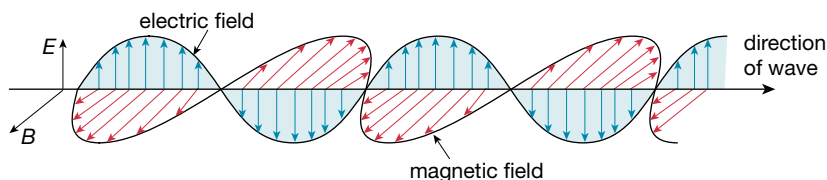


Figure 11.21 The electric and magnetic fields in electromagnetic radiation are perpendicular to each other and are both perpendicular to the direction of propagation of the radiation.

The electromagnetic spectrum

All electromagnetic radiation demonstrates the basic wave properties and, under the right circumstances, all light will reflect, refract, diffract and interfere.

The electromagnetic spectrum is divided into groups on the basis of their method of production. For example, X-rays are produced by the transitions made by the inner electrons of metal atoms, and gamma rays are only emitted from within the nucleus. It is entirely possible for an X-ray to have a higher frequency than a gamma ray. It is just that an X-ray originates from within the electron cloud rather than the nucleus.

A comparison of the wavelength and the size of an object (Table 11.3) is useful when deciding whether an object will reflect or diffract the light. For example, a 3 cm microwave will be reflected from a light aircraft, but a 3 km wave will pass around it as if the plane were not there. This is why microwaves are chosen for use in radar. Similarly, a human cell that is 10^{-6} m across will reflect visible light, but microwaves will just as likely pass through. On the other hand, visible light will not reflect from an atom whose diameter is about 10^{-10} m.

Radio waves

Radio waves were first generated and detected by Heinrich Hertz in 1888 using a method which provided an experimental confirmation for Maxwell's theory. Hertz made sparks jump across a small air gap using a high-voltage alternating current. Each time a spark was created, a burst of radio waves with a frequency of about 10^9 cycles per second (1 GHz) was produced (as predicted). Hertz detected these waves as an induced EMF in a loop of wire located across the room. Essentially, Hertz had built the first radio transmitter and receiver.

By the turn of the century, the Italian inventor Marconi had shown that it was possible to send radio signals across the Atlantic Ocean, and modern telecommunications were

born! Today, radio waves are the basis for many of our communication systems, including AM and FM radio, television, CB radio and cellular telephones.

X-rays

It is only a little over 100 years since the German physicist Wilhelm Roentgen produced the first X-rays. X-rays are created when beams of high-energy electrons are made to collide with metal atoms located in a target. X-rays have very high frequencies and so have correspondingly small wavelengths. Values such as 10^{-12} m are typical.

From the beginning, it was realised that X-rays could be of great value to medicine; Roentgen was able to photograph the bones in his wife's hand as early as 1896. But there was great debate as to whether X-rays were streams of high-energy particles or electromagnetic radiation. Von Laue, a German, showed that X-rays could be diffracted using crystals as a diffraction grating. This indicated that X-rays did indeed have a wave nature, and so they were accepted as part of the electromagnetic spectrum.

Table 11.3 A comparison of electromagnetic wavelengths and everyday objects

Type of wave	Typical wavelength (m)	Comparable object
AM radio wave	~ 100	sports oval
FM radio or TV wave	~ 3	light aircraft
Visible light	$\sim 10^{-7}$	small cell
Ultraviolet	$\sim 10^{-8}$	large molecule
X-ray	$\sim 10^{-10}$	atom
Gamma ray	$\sim 10^{-15}$	atomic nucleus

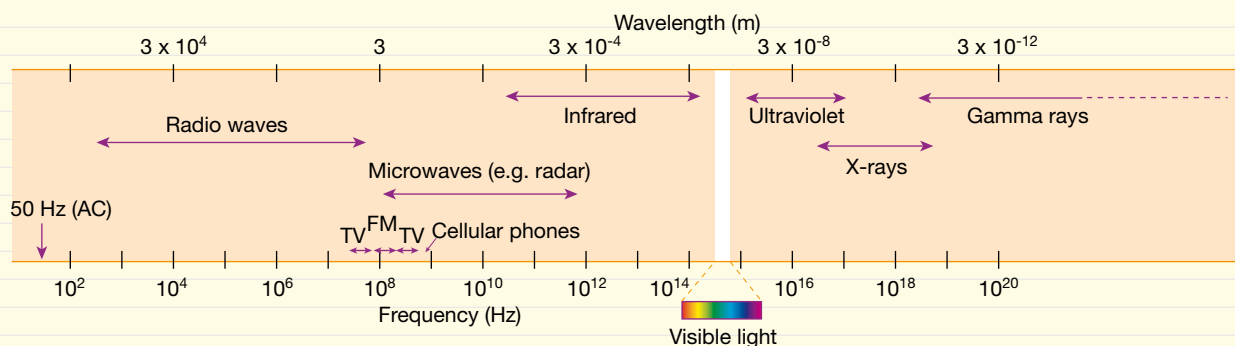


Figure 11.22 The electromagnetic spectrum. Visible light is just one small component of the whole range of possible frequencies within the entire electromagnetic spectrum. As the frequency of the EMR increases, the wavelength decreases. Some radio waves have a wavelength in the order of tens of kilometres while gamma ray wavelengths can be smaller than the diameter of a proton.

William Lawrence Bragg and his father, William Henry Bragg, went further, showing how X-rays could be made to diffract from the various planes of atoms in a crystal. Knowing the X-ray wavelength and the diffraction angles for the X-rays, the distance between atoms and planes of atoms could be determined by using a relatively simple relationship.

X-ray diffraction was very significant in opening up the field of crystallography this century, as it gave physicists and chemists a wonderful tool for probing the atomic arrangement within matter. Crick and Watson were only able to discover the double-helix shape of DNA in the early 1950s by using X-ray diffraction techniques. For their pioneering work, both Braggs—father and son—received a Nobel Prize in 1915.



11.2 summary

The wave model established

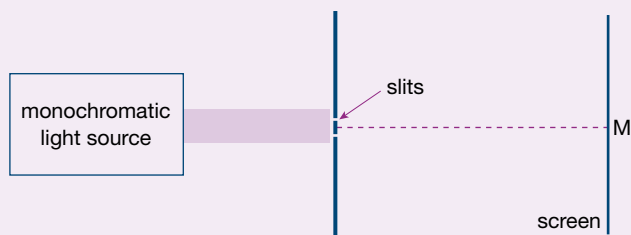
- Waves diffract (bend) as they pass by an obstacle or through a gap.
- The extent of the diffraction is determined by the relative sizes of the wavelength, λ , and the obstacle or gap diameter, w .
- The extent of diffraction of light $\propto \frac{\text{wavelength}}{\text{slit size}} = \frac{\lambda}{w}$.
- The interference pattern seen in a double-slit experiment consists of alternating bright and dark fringes resulting from constructive and destructive interference of monochromatic (single frequency) light. The small wavelength for light requires that the slits are very close to each other and the screen is a significant distance from the slits.
- If the path difference (pd) between the wave trains from the slits $= n\lambda$, $n = 0, 1, 2, \dots$ is met, a bright fringe is seen. If $\text{pd} = (n - \frac{1}{2})\lambda$, $n = 1, 2, 3, \dots$, then a dark fringe is seen.
- If all other factors are held constant, the fringe spacing can be increased by using light of a longer wavelength (lower frequency) or placing the viewing screen or film further from the pair of slits, or decreasing the separation of the slits.



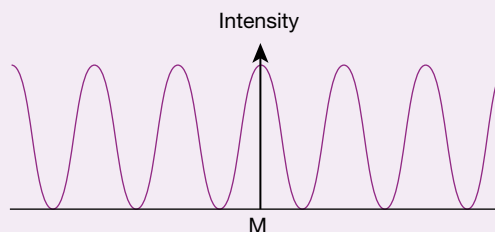
11.2 questions

The wave model established

The following information applies to questions 1–4. A double-slit interference experiment using coherent monochromatic light from a laser is depicted below.



- Why is the intensity a maximum at M?
- Explain how a point of minimum intensity is produced.



- What is a coherent light source?
 - Describe what happens to the laser light as it passes through each slit.
 - The light emerging from each slit is said to be 'in phase'. Explain the meaning of this expression.
- The following diagram shows the resulting intensity pattern after light from the two slits reaches the screen.
 - What will be seen on the screen at:
 - a nodal position?
 - an antinodal position?
- For the interference experiment described in Question 2, what will be seen on the screen at positions where the two wave trains have:
 - a path difference of half a wavelength?



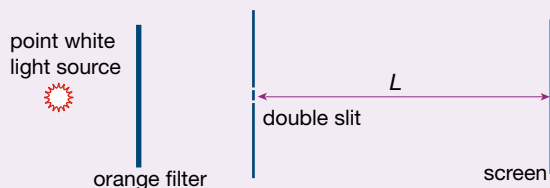
- b** a path difference of one wavelength?
- c** a path difference of one and a half wavelengths?

4 Yellow light is directed through a pair of thin slits separated by $20\text{ }\mu\text{m}$. The slits are exactly 1.0 m from a screen. A diffraction pattern is observed on the screen.

- a** What effect does moving the screen closer have on the observed pattern?
- b** The light source is altered to produce light of a different frequency. The fringe spacing is observed to increase. Was the frequency of light increased or decreased? Justify your choice.

The following information applies to questions 5–8.

A double-slit interference experiment in which an interference pattern is obtained on a screen at a distance L from the slits is depicted below.



- 5** Which one or more of the following changes would result in an increased fringe spacing on the screen?
- A** Replace the orange filter with a red filter.
 - B** Replace the orange filter with a blue filter.
 - C** Increase L .
 - D** Decrease L .
- 6** Which one or more of the following statements is correct?
- A** Increasing the slit separation increases the spacing of the interference fringes.
 - B** Increasing the slit separation decreases the spacing of the interference fringes
 - C** Increasing the brightness of the light source increases the fringe spacing.

- 7 a** If the orange filter is removed and white light passes through the slits, coloured fringes may be observed. What phenomenon is responsible for this?
- b** If the orange filter is put back in place but one slit is blocked, what changes will occur in the pattern?
- 8 a** Two students are trying to replicate Young's double-slit experiment. One uses torch light and the other uses light from a laser. Which student is most likely to obtain the expected diffraction pattern? Why?
- b** If Thomas Young's experiment was modelled using circular waves in a ripple tank, which of the following events would represent constructive interference?
- A** Crests meet troughs.
 - B** Crests meet crests.
 - C** Troughs meet crests.
 - D** Troughs meet troughs.
- 9** Estimate the size of the smallest object that can be clearly imaged by a microscope that uses visible light. Explain this limitation.

11.3 Photoelectric effect: Counterevidence for 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.

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 11.23. 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.

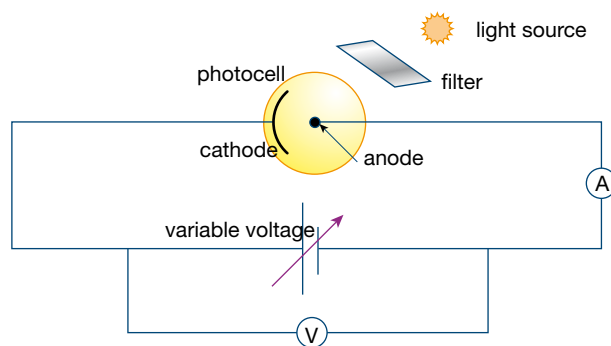


Figure 11.23 In 1916, Robert Millikan used apparatus similar to this to determine the threshold frequency, f_0 , for a variety of metal cathodes. He found that for different metals a different minimum frequency was required in order to produce photoelectrons. He found that elements with higher ionisation 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— $3.00 \times 10^8 \text{ m s}^{-1}$ —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 39

Photoelectric effect



INTERACTIVE TUTORIAL

Photoelectric effect: Investigate colour of light and the PE effect

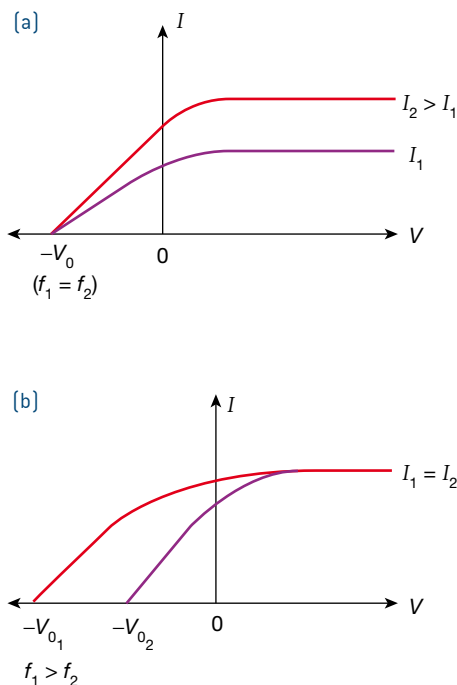


Figure 11.24 Photoelectric current plotted as a function of the applied voltage between the cathode and the anode in the photocell. A standard monochromatic light source where $f > f_0$ shows that with a forward potential, every available photoelectron is included in the current. With a reverse potential, 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 11.24, 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 11.24a). 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 *reverse* 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 opposing 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 voltage) is given by $W = qV$. In this case, the voltage 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 $W = eV_0$. 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.



The photoelectron with the maximum speed has a **KINETIC ENERGY** (J) given by:

$$E_k(\text{max}) = \frac{1}{2}mv_{\text{max}}^2 = eV_0$$

where V_0 = the stopping voltage
 e = the charge on the electron

- 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 (Figure 11.24).

Physics file

Since we are often interested in the speed at which electrons travel, it may be worth rearranging the relationship

$$\frac{1}{2}mv_{\text{max}}^2 = eV_0$$

so that the speed is the subject, i.e.

$$v_{\text{max}} = \sqrt{\frac{2eV_0}{m}}$$

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 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.



1 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 of 100 V, it can automatically be stated that it has gained 100 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 $1 \text{ eV} = qV = 1.6 \times 10^{-19} \times 1 = 1.6 \times 10^{-19} \text{ J}$. Hence:



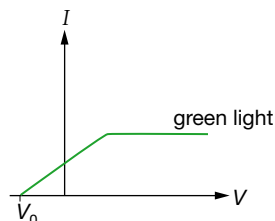
$$1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$$

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 $E_k(\text{max}) = eV_0$. When working in electronvolts, $E_k(\text{max})$ is given directly by the *value* of the stopping voltage in volts. For example, should a stopping voltage of 2.5 V be required, it can automatically be stated that the maximum kinetic energy of any electron is 2.5 eV.

Worked example 11.3A

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.

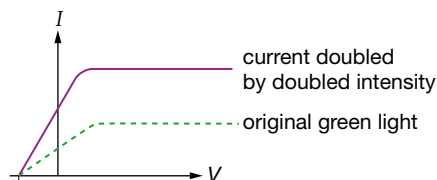


- What current reading would be expected if red light was shone onto the cathode?
- Draw the I - V graph that would result if the intensity of the incident green light was doubled.
- Draw the I - V graph that would result if violet light of a very low intensity was incident upon the cathode.
- When UV 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*.

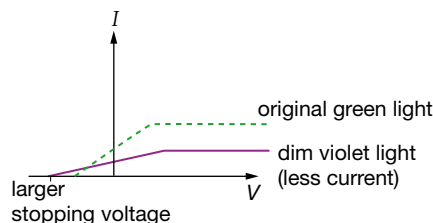
Solution

- Since red light corresponds to light of a lesser frequency than the threshold frequency (yellow light), no current will be observed.

b



c



- Working in joules:

$$\begin{aligned} E_k(\text{max}) &= eV_0 \text{ (joules)} \\ &= 1.6 \times 10^{-19} \times 2.25 \\ &= 3.6 \times 10^{-19} \text{ J} \end{aligned}$$

Working in electronvolts:

$$E_k(\text{max}) = 2.25 \text{ eV (by definition of the electronvolt)}$$

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.



INTERACTIVE TUTORIAL

Photoelectric effect: Investigate forward voltage, reverse voltage

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 $E = Pt$, where P is the power of the beam and t is the time for 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 an 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.

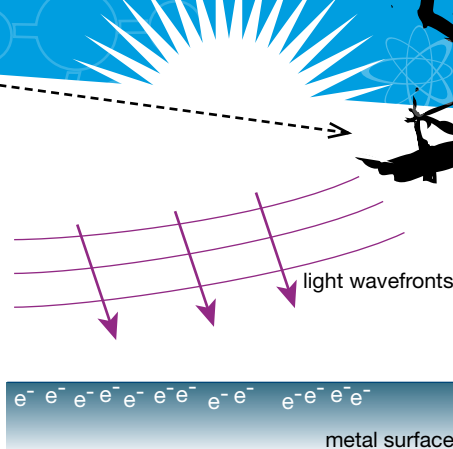


Figure 11.25 The wave model predicts that a wavefront will interact with all electrons near the surface of a material at once. The total energy available will be divided among the electrons. If the light beam has a low intensity, the wave model predicts that there will therefore be a time delay before any electrons are ejected from the metal.

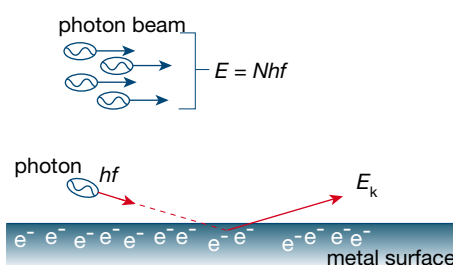


Figure 11.26 The photon model for light suggests that the beam of light carries its energy as a stream of photons, and that each photon interacts with one electron in the metal. A low-intensity beam simply has fewer photons, but each photon still carries the same energy.

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 $E = hf$. Substituting $f = c/\lambda$ gives:

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

where h = Planck's constant = $6.63 \times 10^{-34} \text{ J s}$

f = frequency (Hz)

c = speed of light = $3.00 \times 10^8 \text{ m s}^{-1}$

λ = wavelength (m)

E = 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

Physics file

It is often convenient to give the energy of a photon in the non-SI unit, the *electronvolt*. As $1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$, 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.

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.

Worked example 11.3B

A 100 W light globe produces yellow-green light of wavelength 500 nm. Determine the number of photons released from the globe every minute.

Solution

The globe will release a total energy of $E = Pt = 100 \times 60 = 6000 \text{ J}$ in 1 minute. This energy is carried by N photons, and the energy of each photon will be given by hf , so the total energy will be $E = Nhf$,

and since $f = \frac{c}{\lambda}$, so:

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

Rearranging, this gives:

$$N = \frac{E\lambda}{hc} \\ = \frac{6000 \times 5.0 \times 10^{-7}}{6.63 \times 10^{-34} \times 3.0 \times 10^8} \\ = 1.5 \times 10^{22} \text{ photons emitted each minute—a very large number indeed!}$$



11.3 summary

The photoelectric effect: Counterevidence for wave model

- 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.
- Increasing the forward voltage does not alter the rate of electron release (i.e. the current).
- 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, $E_k(\text{max})$.

$$E_k(\text{max}) = \frac{1}{2}mv^2 = eV_0$$
- The electronvolt is an alternative (non-SI) unit of energy:

$$1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$$
- 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:
 $E_{\text{photon}} = hf$, where $h = \text{Planck's constant} = 6.63 \times 10^{-34} \text{ J s}$ or $4.14 \times 10^{-15} \text{ eV s}$.
- In the photon model, a beam of light consists of a stream of photons. The total energy in the beam will be Nhf , where N is the number of photons in the beam.



11.3 questions

The photoelectric effect: Counterevidence for wave model

For the following questions use:

Planck's constant = $6.63 \times 10^{-34} \text{ J s}$ or $4.14 \times 10^{-15} \text{ eV s}$

speed of light = $3.00 \times 10^8 \text{ m s}^{-1}$

charge on electron = $-1.60 \times 10^{-19} \text{ C}$

mass of electron = $9.11 \times 10^{-31} \text{ kg}$

- 1 The frequency of a green light is $5.60 \times 10^{14} \text{ Hz}$.
 - a What is the wavelength of this light?
 - b Calculate the energy of a photon of this green light in joules and electronvolts.
- 2 Write three key statements that Lenard would have made about the photoelectric effect.
- 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
 - a Convert $3.68 \times 10^{-19} \text{ J}$ into electronvolts.
 - b Convert 2.65 eV into joules.
- 5 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.
- 6 An ultraviolet source produces radiation with a wavelength of $1.5 \times 10^{-8} \text{ m}$.
 - a What is the frequency of this radiation?
 - b Calculate the energy of a photon of this light in joules and electronvolts.
- 7 Draw the I - V graph that would be obtained if a single photocell was first exposed to dim green light and then bright green light of the same frequency. Assume that green light has a frequency above the threshold frequency of the metal cathode.
- 8 The frequency of a red light is $4.0 \times 10^{14} \text{ Hz}$ and the frequency of a gamma ray is $4.0 \times 10^{20} \text{ Hz}$. How many photons of red light are needed to equal the energy of the gamma ray photon?
- 9 An infrared lamp emits light of frequency $1.0 \times 10^{14} \text{ Hz}$. If the lamp has a power output of 60 W, calculate the number of photons emitted by the lamp each second.
- 10 An ultraviolet lamp emits 1.89×10^{20} photons each second. The power output of the lamp is 200 W. Determine the frequency of the radiation being produced.

11.4 The dual nature of light

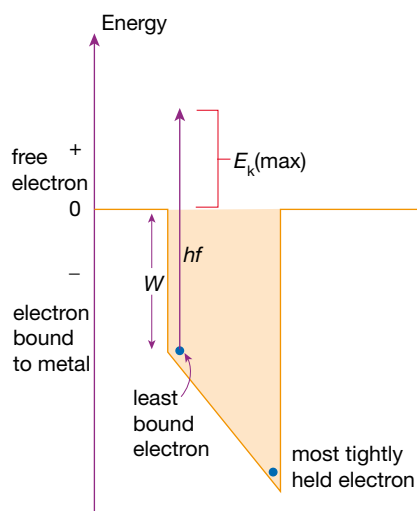


Figure 11.27 Electrons are bound by different amounts of energy. Photons with sufficient energy will free the least bound electron with the greatest speed.

Einstein's explanation of the photoelectric effect

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 = Planck's constant = $6.63 \times 10^{-34} \text{ J s}^{-1}$ or $4.14 \times 10^{-15} \text{ eV s}$
 f_0 = 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, $E_k(\text{max})$. So:

$$hf = W + E_k(\text{max})$$

Rearranging, the maximum kinetic energy for the least bound electron becomes:



$$E_k(\text{max}) = \frac{1}{2}mv_{\text{max}}^2 = hf - W$$

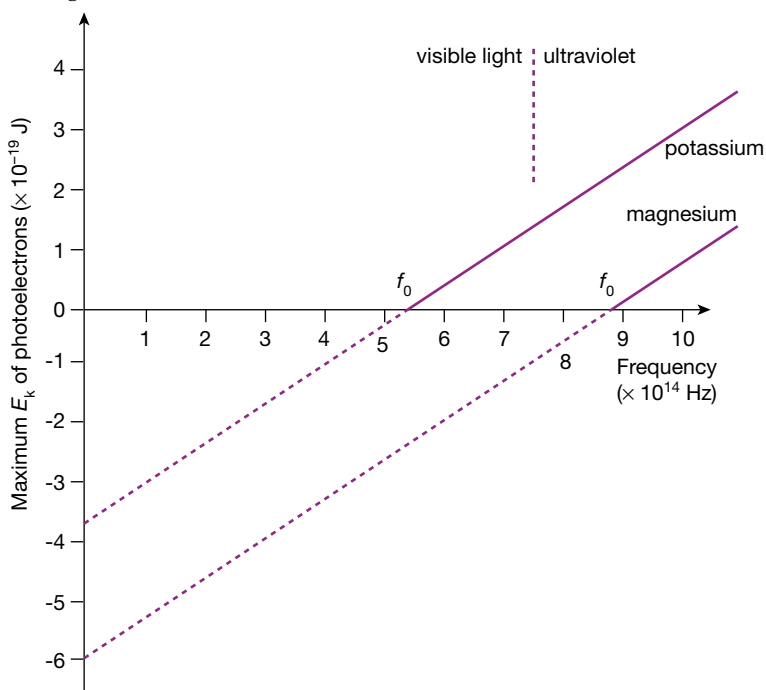
where $E_k(\text{max})$ = the maximum kinetic energy of the released electrons
 hf = the incident photon energy
 W = work function

Note that eV or J may be used as the units of energy.

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:

$$eV_0 = \frac{1}{2}mv_{\max}^2$$

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.



Physics file

By comparison of the form of the equation for a straight line, $y = mx + c$, and Einstein's finding $E_k(\max) = hf - W$, it can be seen that extrapolating the graph of $E_k(\max)$ vs. frequency back to the vertical axis will give the magnitude of the work function, W .

Figure 11.28 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 11.4A

Yellow-green light of wavelength 500 nm shines on a metal whose stopping voltage is found to be 0.80 V. The mass of an electron is 9.11×10^{-31} kg. Find:

- the speed of the fastest moving photoelectron produced
- the work function of the metal in both joules and electronvolts.

Solution

a $E_k(\max) = eV_0 = \frac{1}{2}mv_{\max}^2$

So rearranging:

$$\begin{aligned} v_{\max} &= \sqrt{\frac{2eV_0}{m}} \\ &= \sqrt{\frac{2 \times 1.6 \times 10^{-19} \times 0.80}{9.1 \times 10^{-31}}} = \sqrt{2 \times 8.1 \times 10^{11}} \\ &= 5.3 \times 10^5 \text{ m s}^{-1} \end{aligned}$$

b $E_k(\max) = hf - W$

so, $W = hf - E_k(\max)$

$$\begin{aligned} &= \frac{hc}{\lambda} - eV_0 \\ &= \frac{6.63 \times 10^{-34} \times 3.0 \times 10^8}{5.0 \times 10^{-7}} - 0.80 \times 1.6 \times 10^{-19} \\ &= 3.97 \times 10^{-19} - 1.28 \times 10^{-19} \\ &= 2.69 \times 10^{-19} \text{ J} = 1.68 \text{ eV} \end{aligned}$$

The work function is 2.7×10^{-19} J or 1.7 eV.



INTERACTIVE TUTORIAL

Photoelectric effect: Investigate frequency and maximum E_k

Physics file

When applying the wave equation $v = f\lambda$ to any form of EMR travelling in air, don't forget that $v = c$! Therefore, instead of $f = v/\lambda$, we say $f = c/\lambda$!

Worked example 11.4B

Potassium has a threshold frequency of 5.4×10^{14} Hz, and when it is illuminated by ultraviolet light of frequency 9.0×10^{14} Hz, photoelectrons with a stopping voltage of 1.5 V are produced. Determine a value for Planck's constant in J s.

Solution

$$E_k(\text{max}) = hf - W$$

$$= hf - hf_0$$

$$= h(f - f_0)$$

$$\text{Rearranging, we get } h = \frac{E_k(\text{max})}{f - f_0}$$

$$\begin{aligned} \text{Now, } E_k(\text{max}) &= eV_0, \text{ so, } h = \frac{eV_0}{f - f_0} \\ &= \frac{1.6 \times 10^{-19} \times 1.5}{9.0 \times 10^{14} - 5.4 \times 10^{14}} \\ &= \frac{2.4 \times 10^{-19}}{3.6 \times 10^{14}} \\ &= 6.7 \times 10^{-34} \text{ J s} \end{aligned}$$

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 like 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 $E = hf$ to Einstein's mass-energy equivalent of the photon, $E = mc^2$, i.e. $mc^2 = hf$, hence $mc = hf/c$.

He then interpreted the product ' mc ' to be the **momentum of the photon**, and since $c = f\lambda$, so:

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

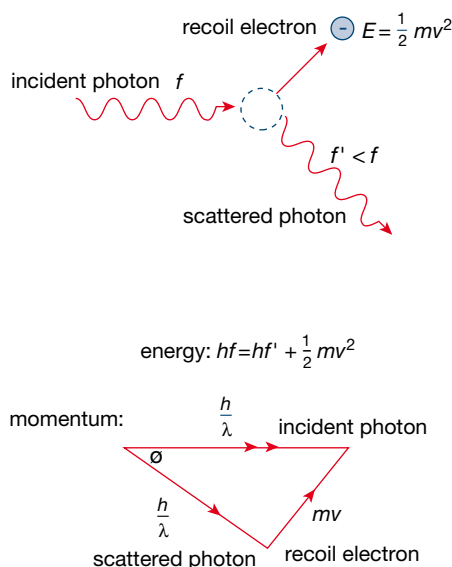


Figure 11.29 Compton scattering involves X-ray photons of a given wavelength giving both energy and momentum to a recoil electron. The collision can be considered in exactly the same way as one might analyse two massive objects interacting. In some situations, the scattering will be elastic, where no 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.



The **MOMENTUM** of a photon, p , is given by:

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

where f and λ are the frequency and wavelength of the radiation respectively. Note that h must be applied in its SI unit, $h = 6.63 \times 10^{-34}$ J s.

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.

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. 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.



11.4 summary

The dual nature of light

- 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.
- The maximum kinetic energy of the photoelectron is determined by the incident photon energy and the work function of the metal according to:

$$E_k(\text{max}) = hf - W = hf - hf_0$$
- By experiment, the maximum kinetic energy for the electrons (i.e. that of the fastest electron) can be found by using a reverse voltage called the stopping voltage, V_0 .
- $E_k(\text{max}) (\text{J}) = eV_0$, where V_0 is the stopping voltage.
- A graph of $E_k(\text{max})$ versus frequency will have a gradient equal to Planck's constant, h , and a 'y intercept' equal to the work function, W .
- The momentum of a photon, p , of wavelength, λ , is given by:
 $p = h/\lambda = hf/c$, where $h = 6.63 \times 10^{-34} \text{ J s}$.
- 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.



11.4 questions

The dual nature of light

For the following questions use:

Planck's constant = $6.63 \times 10^{-34} \text{ J s}$ or $4.14 \times 10^{-15} \text{ eV s}$

speed of light = $3.00 \times 10^8 \text{ m s}^{-1}$

charge on electro = $-1.60 \times 10^{-19} \text{ C}$

mass of electron = $9.11 \times 10^{-31} \text{ kg}$

- 1 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.



- 2 Which one or more of the following properties is light able to display?

A charge
B energy
C mass
D momentum

- 3 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.

- 4 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.

- 5 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.

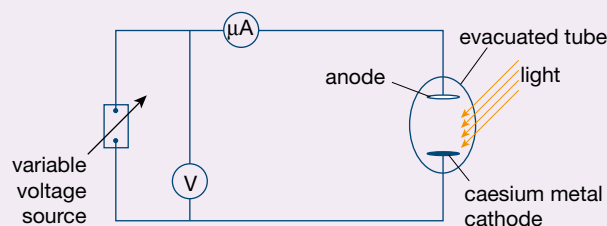
- 6 A student investigating the photoelectric effect obtains the following experimental data when using a caesium cathode.

Category of light	Frequency of light ($\times 10^{14}$ Hz)	Stopping voltage (V)
Blue	7.3	0.9
UV	10.1	2.1

She realises that her data is limited but sufficient to obtain *estimates* of Planck's constant and the work function of caesium. What values will she obtain?

The following information applies to questions 7 and 8.

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.



- 7 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?
- 8 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_0 .
- 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.

chapter review

For the following questions use:

charge on electron:	$e = 1.60 \times 10^{-19} \text{ C}$
mass of electron:	$m_e = 9.11 \times 10^{-31} \text{ kg}$
mass of proton:	$m_p = 1.67 \times 10^{-27} \text{ kg}$
Planck's constant:	$h = 6.63 \times 10^{-34} \text{ J s}$
speed of light:	$c = 3.00 \times 10^8 \text{ m s}^{-1}$

- 1 Explain how the particle approach failed to fully and accurately model the refraction of light.
- 2 A student constructs a set-up identical to Young's double-slit experiment but she uses white light instead of monochromatic light. Fully describe the resulting diffraction pattern.
- 3 X-rays with a photon energy of 40 keV are directed through a sample of aluminium foil and a certain diffraction pattern results. Find the wavelength of the incident X-rays.
- 4 A heavy and a light string are tied together at one end. A progressive wave travelling along the heavy string passes into the lighter string and its velocity increases by 50%. What would you expect to happen to the frequency and wavelength of this wave?

The following information applies to questions 5 and 6.

A student sets up a Young's slit experiment in which yellow light of wavelength 580 nm is incident on two narrow slits whose separation is $1.0 \times 10^{-4} \text{ m}$. An interference pattern is formed at a screen 2.0 m from the slits.

- 5 Describe the pattern seen on the screen.
- 6 Describe the effect on the bright fringe by:
 - i increasing the slit separation
 - ii increasing the distance of the screen from the slits.
- 6 If the slit separation and screen distance are kept constant, describe the effect on the fringe spacing if:
 - a orange light is substituted for the yellow light
 - b blue light is substituted for the yellow light
 - c the brightness of the yellow light source is increased
 - d the brightness of the yellow light source is reduced.
- 7 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?

- 8 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.

- 9 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?

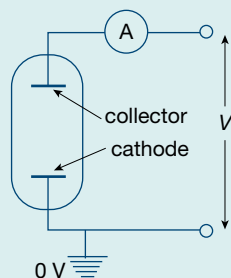
- A It varies with the colour of the incident light.
- B It varies with the intensity of the incident light.
- C It varies with the speed of the incident light.
- D It varies with the incident angle of the light.

- 10 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 11 and 12.

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)	$E_k(\text{max})$ (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

- 11 Plot a graph of the maximum kinetic energy of the ejected photoelectrons (J) as a function of the frequency of the incident radiation (Hz).

- Calculate the gradient of the graph.
- What does the gradient of the graph represent?
- Use the graph to determine the threshold frequency for rubidium.
- Will red light of wavelength 680 nm emit photoelectrons from a rubidium surface? Justify your answer.

- 12 Green light with a frequency of 5.6×10^{14} Hz is directed onto the cathode of the photoelectric cell.

- 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.
- What is the kinetic energy (J) of the fastest photoelectrons emitted from the cathode?
- What is the momentum of the fastest photoelectrons emitted from the cathode?
- What is the magnitude of the minimum retarding potential difference that will prevent photoelectrons reaching the collector of the photoelectric cell?

The nature of matter

This photograph shows the trails left by high-energy sub-atomic 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 have largely replaced cloud chambers for the study of particle collisions. Among other things, they enable us to investigate electron-positron annihilation.

In this chapter, you will tour some of the landmark experiments and theories associated with the development of our understanding of matter.

The development of the bubble chamber and many other devices has allowed some of the mysteries associated with the behaviour of matter to be unravelled. But, as we will see, when physicists find some answers, they invariably reveal new questions as well.

by the end of this chapter

you will have covered material from the study of the nature of matter, including:

- electron diffraction patterns as evidence for the wave-like behaviour of matter
- the de Broglie wavelength
- atomic absorption and emission spectra including vapour lamps
- changes in energy levels of atoms and photon energy, frequency and wavelength
- quantised atom and standing waves as a model of the dual nature of matter.

12.1 Matter waves

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.

In the last section we examined the expression for the *momentum* of a *photon* based upon the value of its wavelength. The expression for momentum was $p = h/\lambda$. De Broglie wondered whether the relationship between the momentum and wavelength of a photon could be a general relationship that applied to photons and particles of matter alike.

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 ($p = mv$) 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 1.5 nm. This electron would be expected to display some behaviour in common with other waves of this same wavelength. A most readily observable behaviour of waves is of course *diffraction*.



Figure 12.1 Louis de Broglie (1892–1987).



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 = the de Broglie wavelength of a particle (m)

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

p = the momentum of the particle (kg m s⁻¹)

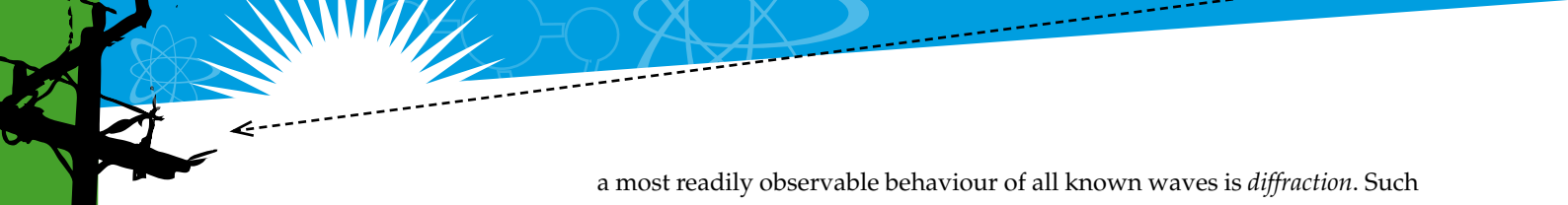
m = the mass of the particle (kg)

v = the velocity of the particle (m s⁻¹)

(Note that the electronvolt value of Planck's constant does not apply when using this formula!)

Matter waves not noticeable every day!

The effects of *de Broglie wavelengths* are not noticeable for matter of everyday speed and mass values, but at the atomic level, this property of matter can be very important when explaining some experimental results. Recall that



a most readily observable behaviour of all known waves is *diffraction*. Such is the strength of the link between diffraction and waves that any entity that is undergoing diffraction would automatically be described as displaying a wave-like property. However, as was discussed earlier, a wave will only undergo *noticeable* diffraction if the size of the wavelength is comparable to the size of the relevant aperture or obstacle.

Although ridiculous, let's calculate the supposed de Broglie wavelength of an everyday item. For example, the de Broglie wavelength of a 150 g cricket ball being bowled with a speed of 30 m s^{-1} would be given by:

$$\lambda = \frac{h}{mv} = \frac{6.63 \times 10^{-34}}{0.150 \times 30} = 1.5 \times 10^{-34} \text{ m}$$

At 10^{-34} m , the de Broglie wavelength associated with the ball is far too small to be detected by diffraction. For it to 'diffract' through a doorway, the opening would need to be comparable to the wavelength, but this is smaller than a tiny fraction of a *proton* diameter (which is 10^{-15} m itself)! Even if the ball were slowed down, its relatively large mass would alone ensure a tiny de Broglie wavelength value and make it impossible for diffraction to be detected.

In other words, the diffraction angle would be so miniscule that the particle in question would just be observed to pass straight through the slit with no wave behaviour observable. This is the reason why in all everyday observations, and indeed all macroscopic experiments, de Broglie wave effects are not observed. In the tiny world of the atom, however, they cannot be ignored!

Given that the expression for the de Broglie wavelength is $\lambda = h/mv$, in order for the de Broglie wavelength of a body to have a large enough value that it may be detected, a very small mass is required. At only $9.1 \times 10^{-31} \text{ kg}$, electrons are just right.

Worked example 12.1A

Find the de Broglie wavelength of an electron that has been accelerated from rest through a potential difference of 50 V. The mass of an electron is $9.11 \times 10^{-31} \text{ kg}$.

Solution

First, find the speed of the electron.

Recall from your Unit 2 studies that the work done on the electron is given by:

$$\begin{aligned} W &= \Delta U_e = qV \\ &= eV \\ &= 1.6 \times 10^{-19} \times 50 \\ &= 8.0 \times 10^{-18} \text{ J} \end{aligned}$$

Therefore, the kinetic energy of the electron after being accelerated through 50 V will be $8.0 \times 10^{-18} \text{ J}$.

So rearranging $E_k = \frac{1}{2}mv^2$:

$$\begin{aligned} v &= \sqrt{\frac{2E_k}{m}} \\ &= \sqrt{\frac{2 \times 8.0 \times 10^{-18}}{9.11 \times 10^{-31}}} \\ &= 4.2 \times 10^6 \text{ m s}^{-1} \end{aligned}$$

Using this velocity to determine the de Broglie wavelength gives:

$$\begin{aligned}\lambda &= \frac{h}{mv} \\ &= \frac{6.63 \times 10^{-34}}{9.11 \times 10^{-31} \times 4.2 \times 10^6} \\ &= 1.7 \times 10^{-10} \text{ m}\end{aligned}$$

This wavelength is still tiny, but note that it is much larger than the de Broglie wavelength of the cricket ball discussed earlier.

The wavelength of the electron's matter wave in Worked example 12.1A is of the order of 10^{-10} m, which is comparable to the spacing between the atoms in a crystal. It is conceivable, therefore, that a beam of these electrons might produce a diffraction pattern after passing through the gap between atoms in a crystal. This is exactly what was eventually done in order to prove de Broglie right.

Physics in action

X-ray scattering inspires quest to find electron diffraction

Before X-rays were *passed through* the gap between atoms in a crystal, a slightly different and very important X-ray interference pattern was observed. In an attempt to find out what X-rays actually were, the choice of a crystal structure to investigate the possible diffraction of X-rays was quite logical to physicists. The spacing of atoms in a crystal is regular and lattice-like, with the layers in a crystal being less than 1 nm apart. Suppose that a crystal of sodium chloride is used. The uniformly spaced atoms are shown in Figure 12.2. A beam of X-rays shone into the crystal will be scattered from the top layer and the next etc.

The X-rays reflecting from the deeper layer have travelled a slightly greater distance than the others. The X-rays emerging from the crystal will interfere with each other, just like the light waves passing through the two slits in Young's famous double-slit experiment or the diffraction gratings studied earlier. This phenomenon is described as *interference or scattering*. There will be locations, determined by path difference values, where constructive interference occurs and locations where destructive interference occurs. These result in the bright and dark bands of the diffraction patterns displayed in section 11.2. It was as early as 1913 that this mode of diffraction of X-rays by crystals had been analysed by W. H. and W. L. Bragg.

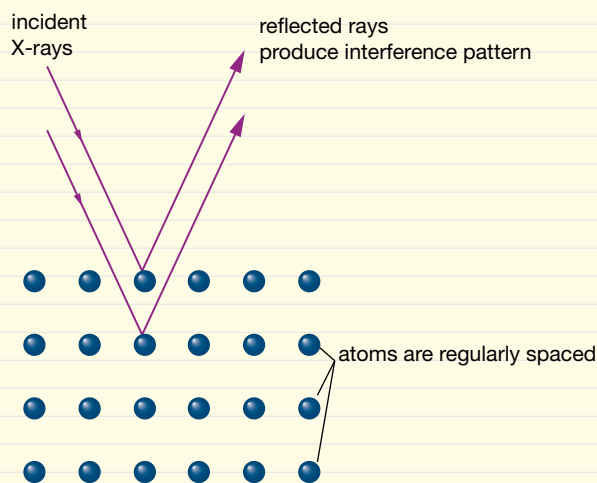


Figure 12.2 The X-rays reflecting from different layers within the crystal structure were known to create an interference pattern like that produced by diffraction gratings.

Matter-scattering experiments

The experimental evidence that validated de Broglie's ideas came in 1927. Davisson and Germer bombarded nickel crystals with beams of electrons, and measured the intensity of the electrons as they scattered from the nickel in the different directions with their movable detector. A diagram of their apparatus is shown in Figure 12.3. The electron gun provides a collimated beam of electrons whose speed is known, because they have been accelerated through a known voltage.

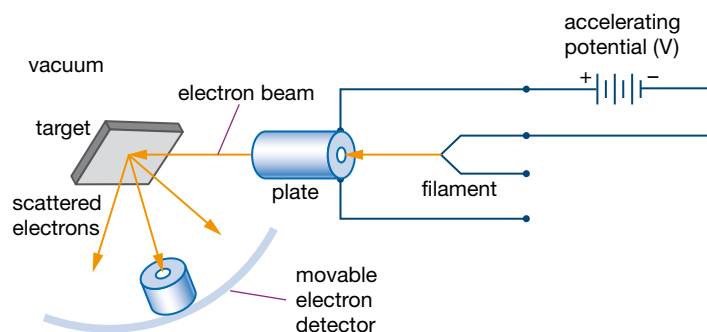


Figure 12.3 The Davisson and Germer apparatus to show electron scattering.

Synchrotron X-rays are an ideal tool for X-ray crystallography. For more information see Chapter 13 'Synchrotron and applications'.

Physics file

Bragg's equation describes the location of maxima in the diffraction pattern produced when X-rays are diffracted by a crystal:

$$n\lambda = 2d\sin\theta$$

If X-rays of a known wavelength, λ , are used, the resulting scattering angles, θ , tell us the distance between atoms, d .

This technique, called X-ray crystallography, has been used to determine the structures of items as complex as DNA.

The detector could be swung around on an axis so that it could intercept electrons scattered from the nickel target in any direction in the plane shown. Davisson and Germer found that as they moved their detector through the different scattering angles, they encountered a sequence of maximum and minimum intensities. They were confident that *diffraction* of the electron beam had occurred. This situation was similar to the scenario of X-ray scattering discussed in the Physics in action on page 439. The electrons were being scattered by the different layers within the crystal lattice. Electrons emerging from the different layers within the lattice had undergone the *interference* that is so firmly established as a wave property.

Recall our study of Young's double-slit experiment. In our analysis of it, the *spacing* of the fringes of the diffraction pattern was used to determine the wavelength of the 'wave' being diffracted. A process can be carried out for this set-up which is very similar to the X-ray diffraction that had been analysed earlier by Bragg (see adjacent Physics file). Davisson and Germer used the spacing of their diffraction pattern to arrive at an experimental value of the apparent matter wavelength, λ , of the electron. The diffraction pattern indicated that the wavelength must have been 0.14 nm.

They then checked out de Broglie's hypothesis. Since the accelerating potential of the electron gun was known to be 75 V, the de Broglie wavelength could be obtained by using exactly the same process as shown in Worked example 12.1A.

First, the speed of the electron is determined. Recall that the work done on the electron is given by:

$$\begin{aligned} W &= \Delta U_e = qV \\ &= 1.6 \times 10^{-19} \times 75 \\ &= 1.2 \times 10^{-17} \text{ J} \\ &= \text{the kinetic energy of the electron} \end{aligned}$$

$$\begin{aligned} \text{And so: } v &= \sqrt{\frac{2E_k}{m}} \\ &= \sqrt{\frac{2 \times 1.2 \times 10^{-17}}{9.11 \times 10^{-31}}} \\ &= 5.1 \times 10^6 \text{ m s}^{-1} \end{aligned}$$

Knowing that the mass of an electron is 9.11×10^{-31} kg and using this velocity to determine the de Broglie wavelength gives:

$$\begin{aligned}\lambda &= \frac{h}{mv} \\ &= \frac{6.63 \times 10^{-34}}{9.11 \times 10^{-31} \times 5.1 \times 10^6} \\ &= 1.4 \times 10^{-10} \text{ m} \\ &= 0.14 \text{ nm}\end{aligned}$$

De Broglie's hypothesis had been verified! Particles of matter had displayed wave-like properties (i.e. diffraction patterns) and furthermore their matter wavelength could actually be calculated by using his equation! De Broglie's formula predicted a value for a particle's wavelength, and the particle was shown to diffract just as if it was a wave of this exact wavelength.

In the same year that Davisson and Germer conducted their experiment, other supporting evidence was forthcoming from G. P. (son of J. J.) Thomson. Rather than scatter an electron beam from a crystal, Thomson produced a diffraction pattern by passing a beam of electrons *through* a tiny crystal. Thomson then repeated his experiment, using *X-rays of the same wavelength* in place of the electrons. The X-ray diffraction pattern was identical to the one made with electrons. De Broglie had been vindicated, and the symmetry in nature he had believed in was a reality.

A 'powder diffraction' technique is commonly adopted for crystal diffraction. Rather than use pure large crystal samples that are difficult to obtain, physicists grind small crystals into a powder so that it can be assumed that it contains numerous tiny crystals of random orientation. As the electron beam is shone onto the powder the diffraction will happen in all three planes, resulting in the *circular* diffraction patterns shown in Figure 12.4.

Synchrotron radiation is ideal for use in powder diffraction techniques. This is utilised in the identification of compounds with a particularly small grain size. For more information see Chapter 13 'Synchrotron and applications'.

Comparing the wavelengths of photons and matter

The de Broglie hypothesis was further verified as all around the world physicists used various particles and observed their diffraction by crystals of various types. Since then hydrogen atoms, helium atoms, protons and neutrons, for example, have all been diffracted and the de Broglie relation has held for all of them. The inherent *wave* nature of matter has been established.

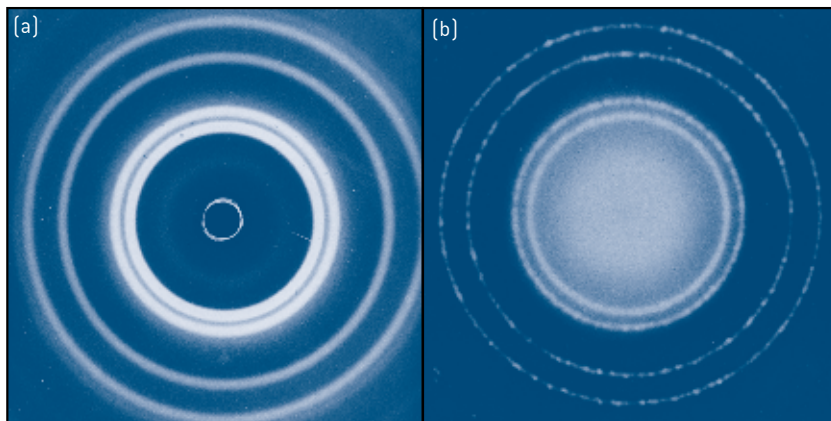


Figure 12.4 These diffraction patterns were taken by using (a) X-rays and (b) a beam of electrons with the same target crystal. Their similarity suggests a wave-like behaviour for the electrons.

Worked example 12.1B

Figure 12.4 shows two images that have been obtained by scattering X-rays and electrons off a sample of many tiny crystals with random orientation. Assume that the X-rays used had a frequency of 8.6×10^{18} Hz and $c = 3.0 \times 10^8$ m s⁻¹.

- Explain why a 'circular' diffraction pattern is obtained, rather than a row of diffraction 'fringes' as studied earlier.
- Since the two diffraction patterns show exactly the same fringe spacing, what conclusions can be made about the electrons and X-rays that were used.
- Determine the de Broglie wavelength of the electrons.
- Compare the momentum of one electron to the momentum of one X-ray photon.

Solution

- The 'gap' through which the X-rays and electrons are passing is not a vertical slit as studied earlier. The gaps between atoms in the crystals will have a range of orientations, resulting in circles rather than fringe lines.
- Since the same crystals were used in each case, we can assume that the electrons and X-ray photons were passing through 'gaps' of the same sizes, resulting in the similar diffraction pattern. Since, for a set gap size, the spacing of the fringes (i.e. the extent of diffraction) is dependent upon the *wavelength* of the photon or particle passing through the crystal, we can conclude that the electrons 'have the same wavelength as the X-rays'. The wavelengths of particles are referred to as 'de Broglie wavelengths'.
- $$\begin{aligned}\lambda_{\text{electrons}} &= \lambda_{\text{X-ray photons}} \\ &= \frac{c}{f} \\ &= \frac{3.0 \times 10^8}{8.6 \times 10^{18}} \\ &= 3.5 \times 10^{-11} \text{ m}\end{aligned}$$
- For both photons and particles: $p = \frac{h}{\lambda}$
Since the X-rays and electrons have equivalent 'wavelengths', they must have equivalent momentum.

Physics in action

Electron microscopy

When viewing small objects through ordinary optical microscopes, we are limited by the diffraction of light. The result is that small bodies become fuzzy. One solution is to use smaller wavelengths, and UV light can double the resolving power of a microscope. X-ray microscopes, however, are out of the question since there are no lenses possible for this type of light. However, *electron beams* can be focused by both electric and magnetic fields (a lens equivalent), and the wavelength of the beam can be adjusted at will by varying the accelerating voltage of the electron gun.

As early as 1937, an electron microscope was built with a magnification of 7000 \times (compared with 2000 \times for the best optical microscope). In an optical microscope, the lenses create a single plane of focus, but a significant advantage of the electron microscope is that the whole object remains in focus. Today, scanning electron microscopes are commonplace in research laboratories, and magnifications to 100 000 \times with a resolution of 10 nm are typical. A scanning tunnelling electron microscope can resolve points closer than 0.1 nm, so that for the first time, pictures of the electron distribution around a molecule are possible.

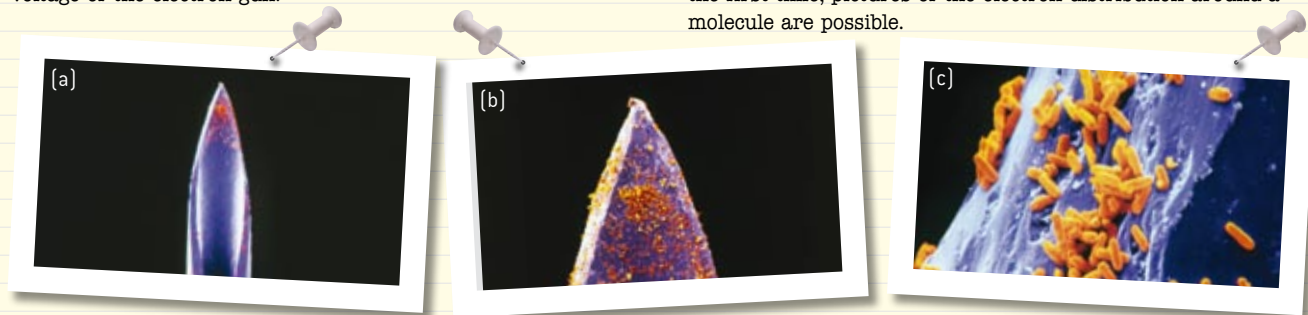


Figure 12.5 Rod-shaped bacteria (orange) clustered on the point of a syringe used to administer injections. The magnifications are (left to right) $\times 9$, $\times 36$ and $\times 560$ at 35 mm size.

Thomson's e/m experiment

Our knowledge and use of properties of electrons are only relatively recent accomplishments in science. It was not until 1897 that physicists were able to shed any light on the internal physical structure of the atom. In that year, Joseph John Thomson demonstrated that cathode rays—rays emanating from a heated cathode in a vacuum—were particles that are fundamental constituents of every atom. For the first time, the atom was shown to have component particles; perhaps, after all, it was not indivisible. To indicate their importance, cathode rays were renamed *electrons*. In 1919 Lord Rutherford identified the proton, and the neutron was discovered by Chadwick in 1934.

Thomson's experiment with cathode rays was performed in two stages. At first the forces on a beam of electrons were balanced using an electric and a magnetic field. This enabled Thomson to find the speed of the electrons. Then the magnetic field was switched off, and the beam was deflected under the influence of the electric field alone. The deflection of the beam was measured, allowing him to find the charge-to-mass ratio (e/m) for the cathode rays. Thomson repeated the experiment with a variety of different cathodes to show that all cathode rays yielded the same value. His result produced a value of about $1 \times 10^{11} \text{ C kg}^{-1}$; the accepted value today is $1.76 \times 10^{11} \text{ C kg}^{-1}$.



Figure 12.6 J. J. Thomson at Cambridge in the late 1890s. The laboratories in which Thomson worked became a centre for inspired research into the atom. Between 1897 and 1934, the electron, nucleus, proton and neutron were discovered there.

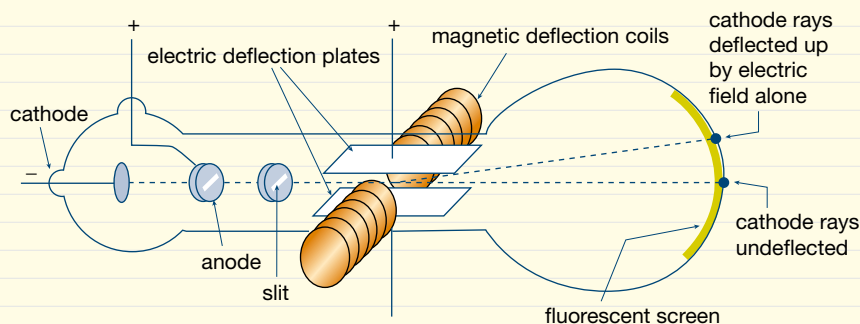


Figure 12.7 J. J. Thomson's apparatus for finding the charge-to-mass ratio (e/m) for cathode rays. In 1897, Thomson used an electron gun to produce a beam of electrons that could be deflected by an electric and magnetic field within an evacuated tube.

The charge on the electron

Although J.J. Thomson was able to show that electrons were components of every atom, he was only able to determine the charge-to-mass ratio for the electron—not its charge or mass alone. These values were not identified until an American, Robert Millikan, was able to determine the electronic charge in 1909.

Millikan measured the total charge on thousands of small oil droplets, since he reasoned that then the charge on the electron must equal the smallest difference in charges on the droplets. Furthermore, this value ought to divide into each charge, yielding an integral number.

Millikan introduced a fine spray of oil drops into the region between two charged plates. The act of creating the spray caused the individual droplets to become charged. Either some electrons were knocked off or some were added through friction. Once in the experimental chamber, the drops were illuminated and viewed through a microscope. Millikan could pick out a droplet and alter the potential difference between the plates until its weight was balanced by the electric force, i.e. $F_g = F_e$. Substituting for these forces, we obtain $mg = Eq$, so $q = mg/E$.

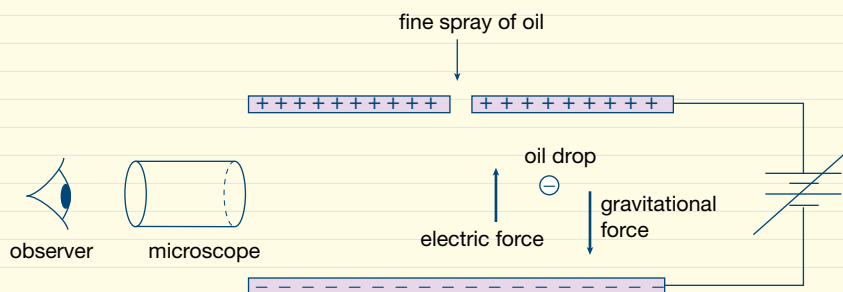


Figure 12.8 A schematic diagram of the oil drop experiment performed by Robert Millikan in 1908 and 1909. As they left the atomiser, the droplets of oil became charged, but in subsequent experiments Millikan used X-rays to charge the oil drops.

Millikan could find the mass by allowing the droplet to fall in air. In doing so, terminal velocity is quickly reached, and Stokes's law allows for the mass to be determined.

Millikan's results suggested a value of 1.63×10^{-19} C for the electron, which compares favourably with the accepted value today of 1.6022×10^{-19} C. Since then, no-one has found

a charged particle with a charge other than a multiple of this value. For this reason, we say that charge is *quantised*—the total charge, q , on a body will be a multiple of the charge e on one electron; that is, $q = ne$.

Another important result from Millikan's experiment was an estimate of the mass of the electron, 9.11×10^{-31} kg—a very small number indeed.

The cathode ray tube is a simple type of particular accelerator. See Chapter 13 'Synchrotron and applications'.

In this chapter we have seen how the relationship between electromagnetic waves and photons has been paralleled by the relationship between de Broglie matter waves and particles. In circumstances where the wave behaviour of matter becomes significant (i.e. when dealing with tiny particles) even classical wave mechanics has been replaced by *quantum mechanics* (in response to more and more sophisticated observations regarding the behaviour of light and matter).

In such circumstances, it is acknowledged that the path of any particular particle cannot be predicted, any more than we could trace the path of a particular photon! Rather, under the theories of *quantum mechanics*, we look at the *probability* that particles or photons will take particular paths. The interference pattern of the type that we have been examining is interpreted as being the *probability distribution* of a matter wave. That is, a mathematical wave function is stated that gives the probability of an electron striking each point on the screen. An understanding of these functions is outside the requirements of this course. In Figure 12.9 the yellow dots in the photographs represent electrons after they have passed through a Young's double-slit apparatus. With only a few electrons, the pattern appears random, but with more and more electrons, the interference fringes are evident. The pattern can only be interpreted as the distribution of the probability of electrons reaching each point on the screen.

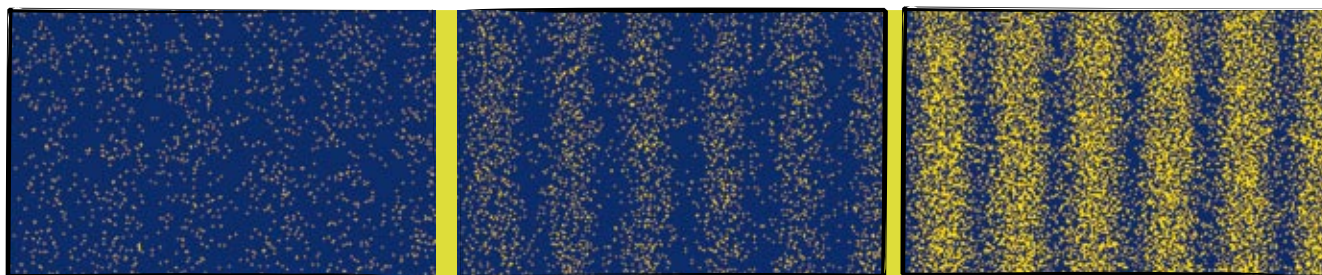


Figure 12.9 Individual electron paths are not predictable, but the physics theories that propose a probability distribution for the diffracted electrons elegantly model the experimental results.

So, there it is: light and matter each have a dual nature, and the problem is not so much that they cannot be reconciled, but that neither light nor matter conforms to a simple model. Our observations of light and matter have led us to quantum theory, which, in turn, as the next section shows, reveals a great deal about the inside of the atom.



12.1 summary

Matter waves

- The de Broglie wavelength, or wavelength of the matter wave, of a body of mass m , moving with a speed v is given by:

$$\lambda = \frac{h}{p} = \frac{h}{mv}$$

- De Broglie suggested his equation by using ideas of symmetry in nature, and scattering experiments provided evidence for the wavelike nature of matter.
- In particle-scattering experiments, beams of particles (electrons usually) are made to travel with a speed so that their matter wavelength approximates the interatomic spacing in a crystal. Consequently, a

diffraction pattern is produced which can only be explained if matter has a wave nature.

- If photons and matter particles being scattered by the same crystal sample produce the same fringe spacing, then they must have the same wavelength and momentum.
- All matter, like light, has a dual nature. Through everyday experience matter is particle-like, and under some situations it has a wavelike nature. This symmetry in nature—the dual nature of light and matter—is referred to as wave-particle duality.



12.1 questions

Matter waves

For the following questions use:

Planck's constant = $6.63 \times 10^{-34} \text{ J s}$

speed of light = $3.00 \times 10^8 \text{ m s}^{-1}$

mass of proton = $1.67 \times 10^{-27} \text{ kg}$

mass of neutron = $1.67 \times 10^{-27} \text{ kg}$

mass of electron = $9.11 \times 10^{-31} \text{ kg}$

mass of α particle = $6.67 \times 10^{-27} \text{ kg}$

- 1 Louis de Broglie's investigation into the existence of matter waves predicts that:

- A all particles exhibit wave behaviour
- B only moving particles exhibit wave behaviour
- C only charged particles exhibit wave behaviour
- D only moving charged particles exhibit wave behaviour.

- 2 Calculate the de Broglie wavelength for the following:

- a a high-speed electron travelling at $1.00 \times 10^7 \text{ m s}^{-1}$
- b a proton with momentum $1.67 \times 10^{-21} \text{ kg m s}^{-1}$
- c a neutron with kinetic energy $3.00 \times 10^{-18} \text{ J}$
- d a 400 g baseball travelling at 10 m s^{-1} .

- 3 a What is the de Broglie wavelength of a 40 g bullet travelling at $1.0 \times 10^3 \text{ m s}^{-1}$?

b Explain why it is impossible for this bullet to exhibit any diffraction effects.

c Are there any circumstances in which objects normally encountered in everyday life could exhibit a detectable de Broglie wavelength? Justify your answer.

- 4 At what speed would a proton travel if it were to have the same wavelength as:

a a gamma ray of energy $6.63 \times 10^{-14} \text{ J}$?

b a photon of yellow light of energy 2.15 eV?

c an electron with momentum $1.82 \times 10^{-24} \text{ kg m s}^{-1}$?

- 5 An electron is accelerated from rest in a uniform electric field through a potential difference of 50.0 V.

a What is the gain in kinetic energy of the electron?

b What is momentum of the electron?

c Calculate the de Broglie wavelength of the electron.

- 6 A charge q with mass m is accelerated from rest through a potential difference ΔV . Derive an expression that defines its de Broglie wavelength, λ , in terms of m , q and ΔV .

- 7 A crystal of Gunners-Love alloy has an interatomic spacing of $1.0 \times 10^{-10} \text{ m}$. Which of the following would be most likely to undergo appreciable diffraction when fired at a thin layer of the crystal? Justify your answer with a mathematical calculation.

A 50 eV neutrons

B 100 eV protons

C 1.0 keV electrons

D X-rays of frequency $1.0 \times 10^{18} \text{ Hz}$

- 8 What accelerating potential difference would be required to give each of the following particles a de Broglie wavelength of 2.0 nm?

a an electron

b a proton

c an alpha particle

12.2

Photons shed light on atom structure

In section 11.3 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

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.

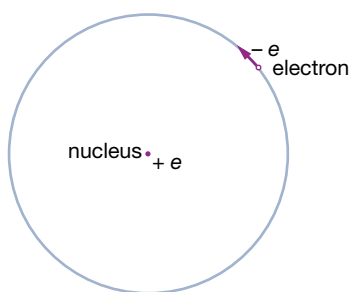


Figure 12.10 Rutherford's nuclear model of the hydrogen atom of 1911 didn't explain the atom's known stability.



PRACTICAL ACTIVITY 40

Observing emission spectra

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 12.11). The emitted light could be divided up into its different component frequencies by a spectrograph (essentially the diffraction gratings studied in section 11.2). Each type of atom was found to be capable of emitting a unique set of frequencies, its *emission spectrum* (Figure 12.12). Thus the emission spectrum of an atom became an individualised property of the atom.

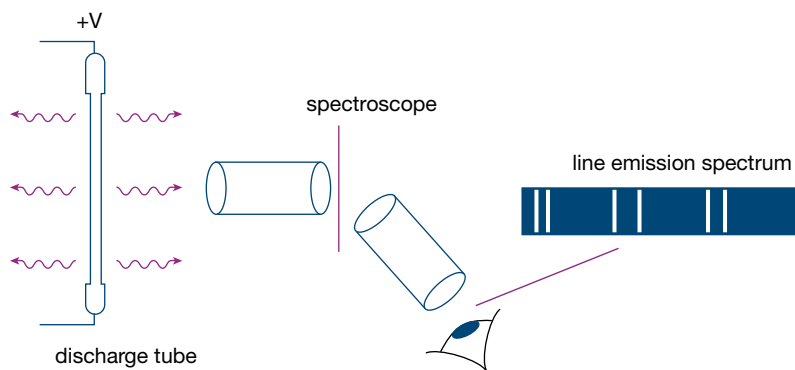


Figure 12.11 Placing gas into a tube and applying a high voltage to it is one technique for producing an emission spectrum. Sodium and mercury are two common examples of metal-vapour lamps.

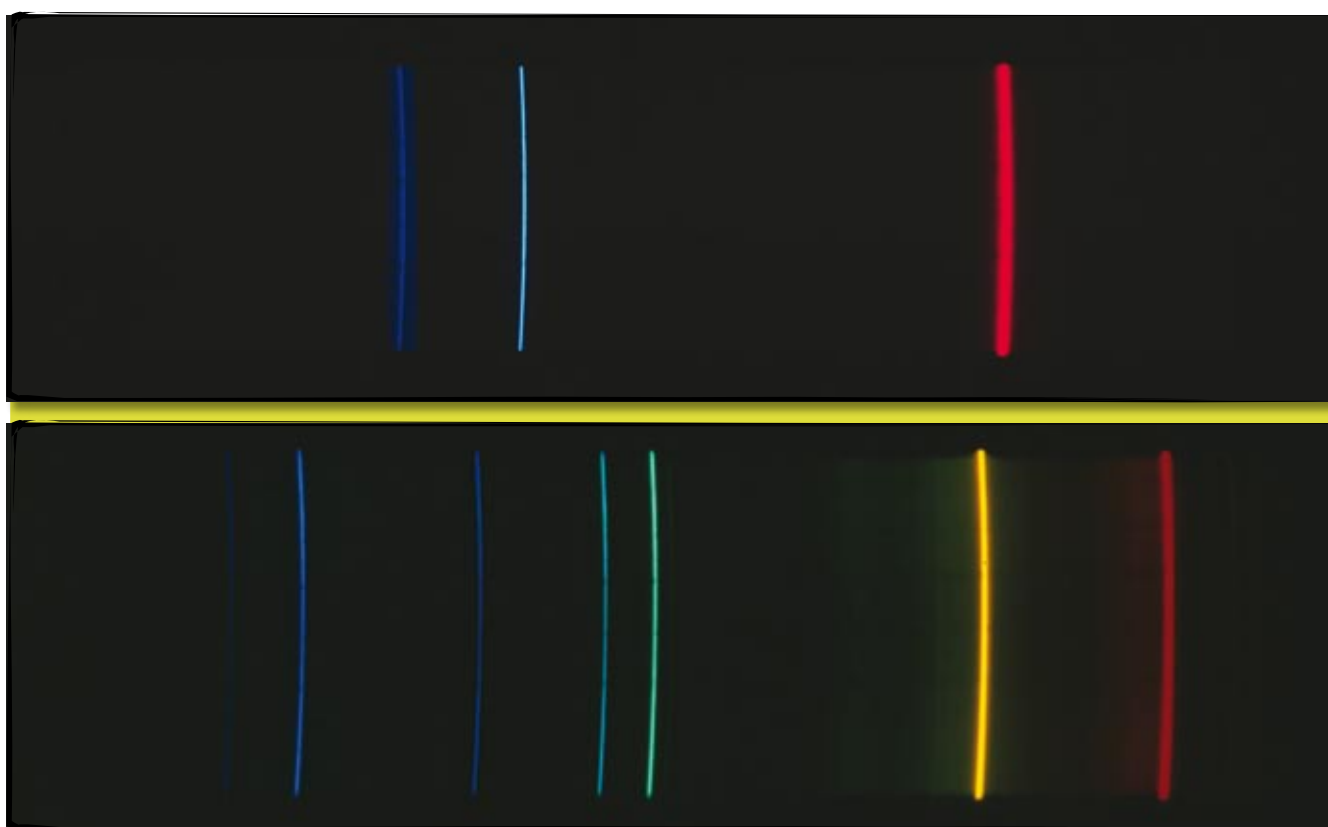


Figure 12.12 The emission spectra of hydrogen and helium, showing only the principal lines.

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 = the photon energy (J or eV)

h = Planck's constant = 6.63×10^{-34} J s or 4.14×10^{-15} eV s

c = the speed of light = 3.0×10^8 m s $^{-1}$

λ = the wavelength (m)

Hydrogen atoms produce the simplest spectrum. Figure 12.12 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 couldn't *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

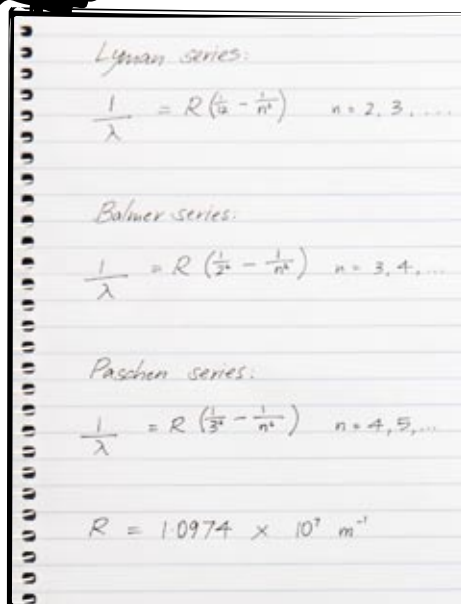


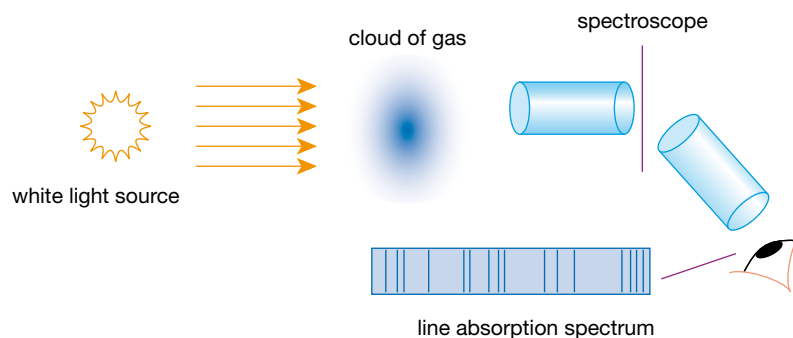
Figure 12.13 Years before emission spectra were understood, mathematical descriptions of the series of frequency values for hydrogen were devised.



PRACTICAL ACTIVITY 41

The spectrometer

Figure 12.14 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’.)

expressions described the three sets of frequencies emitted by hydrogen (see Figure 12.13). 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.

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 12.14. 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.

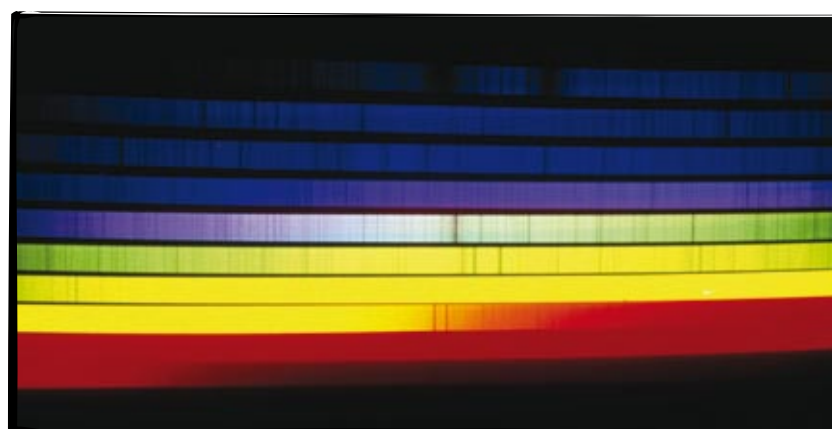


Figure 12.15 The solar spectrum as seen at the Earth. The dark absorption lines in the spectrum are caused by particular frequencies of light being absorbed by cool gases surrounding the Sun.

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 (discussed on page 455). 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 for 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 $n = 1, 2, 3, \dots$ *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.

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 12.16 (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.



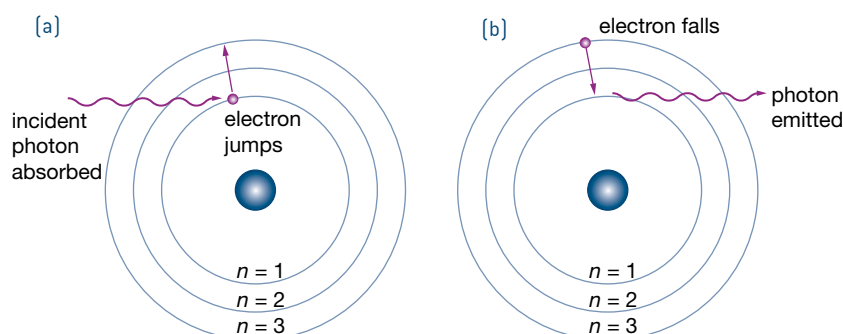
If an electron in an atom moves between energy levels m and n , the energy of the photon that is either emitted or absorbed is given by:

$$E_{\text{photon}} = hf = E_m - E_n$$

where h = Planck's constant = $6.63 \times 10^{-34} \text{ J s}$

f = 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.



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 carrying just the right amount of energy 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 12.3).

Energy level diagrams

Energy level diagrams can be thought of as a section of the electron orbit diagrams shown in Figure 12.16, and a sample is shown in Worked example 12.2A. The allowed energy levels must be stated and these are written along

Physics file

In his explanation of atomic spectra, Bohr only allowed electrons to sit in orbits that satisfied the equation:

$$mvr = \frac{nh}{2\pi}$$

where n is the energy level number.

The product of the mass, velocity and radius, 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.

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 12.2A

The energy levels for atomic mercury are depicted in the diagram.

- 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?
- Determine the wavelength of the light emitted after an electron in an excited mercury atom makes the transition from $n = 3$ to $n = 1$.

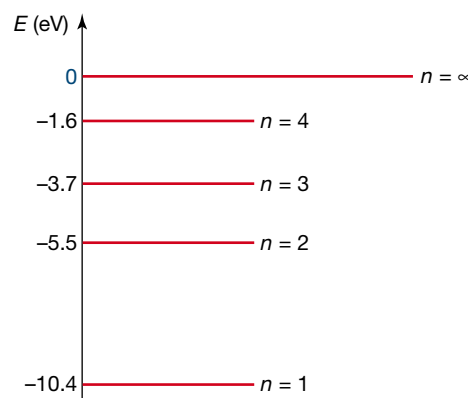
Solution

- 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.

$$\begin{aligned} \text{b } \Delta E &= |E_m - E_n| \\ &= hf \\ &= \frac{hc}{\lambda} \end{aligned}$$

Rearranging gives:

$$\begin{aligned} \lambda &= \frac{hc}{\Delta E} \\ &= \frac{6.6 \times 1.6 \times 10^{-34} \times 3 \times 10^8}{(10.4 - 3.7) \times 1.6 \times 10^{-19}} \\ &= 1.85 \times 10^{-7} \text{ m} \\ &= 185 \text{ nm (in the ultraviolet)} \end{aligned}$$



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.

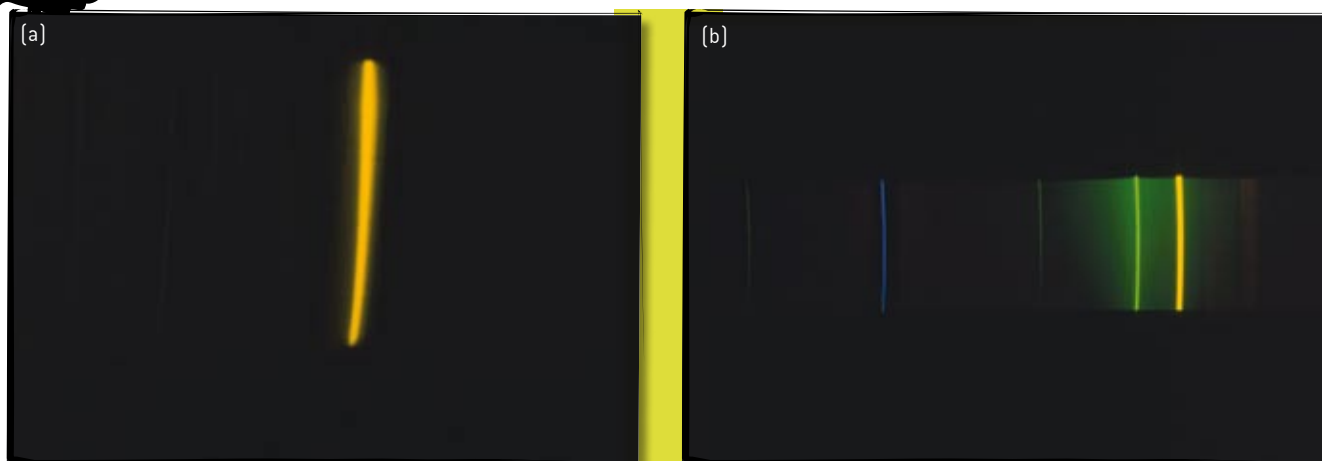


Figure 12.17 Emission spectra of (a) sodium and (b) mercury. The emission spectrum of a sodium- or mercury-vapour lamp consists of a large number of discrete frequencies, some of which lie in the visible region of the spectrum. Note the yellow line that gives sodium light its characteristic colour. Mercury's colours are blue, cyan, green and yellow.

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.

Two common metal-vapour lamps use sodium and mercury. A sodium-vapour lamp generates line radiation at 589.0 nm 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.



12.2 summary

Photons shed light on atom structure

- 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.
- 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:

$$E_{\text{photon}} = hf = E_m - E_n$$

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.



12.2 questions

Photons shed light on atom structure

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 questions 1 and 2.

- _____ ionisation level
- _____ $n = 3$ second excited state
- _____ $n = 2$ first excited state
- _____ $n = 1$ ground state

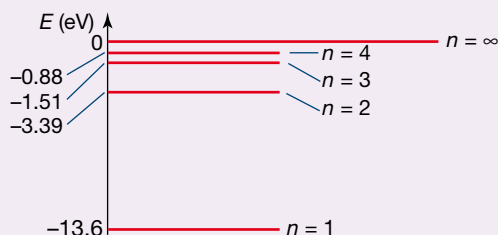
1 A particular atom has four energy levels as shown in the diagram. Explain the meaning of the following terms:

- a** quantisation
- b** ground state
- c** excited states
- d** ionisation energy.

2 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 3–5.



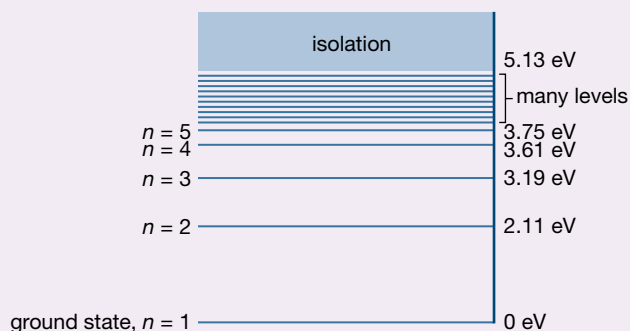
- 3 a** Calculate the frequency of photons emitted when electrons in atomic hydrogen make the transition from the first excited state to the ground state.
- b** What is the wavelength of light that is emitted when the electron in atomic hydrogen goes from the second excited state to the ground state?
- c** What is the minimum energy (in eV) required to ionise a hydrogen atom from the ground state?
- 4 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?
- 5** 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.
- 6** 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.
- 7** Use Bohr's model of the atom to explain why hydrogen was capable of emitting quanta of the exact energy value that it was able to absorb.

The following information applies to questions 8–10.

The diagram shows the energy levels of a sodium ion.



- 8** A sodium atom is in an $n = 3$ excited state. Calculate the shortest wavelength of light that this atom could emit.
- 9** Numerous excited sodium atoms have electrons in the $n = 5$ level. How many different photon energies may be observed as these atoms de-excite?
- 10** Are sodium atoms able to emit or absorb more frequencies of light? Explain.

12.3

Bohr, de Broglie and standing waves

In the previous section it was shown that Bohr's model of the atom could account for many of the observations made about absorption and emission spectra. All that remained to be done was to check whether it would account for the actual energies of the hydrogen spectra, or indeed the spectra of other elements.

Bohr's energy levels of hydrogen

Figure 12.18 shows the measured energy levels for hydrogen with its ionisation energy 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 = the energy of the n th level for hydrogen (eV)

n = the energy level number 1, 2, 3, ...

The negative sign indicates that energy must be added to the system to excite the atom.

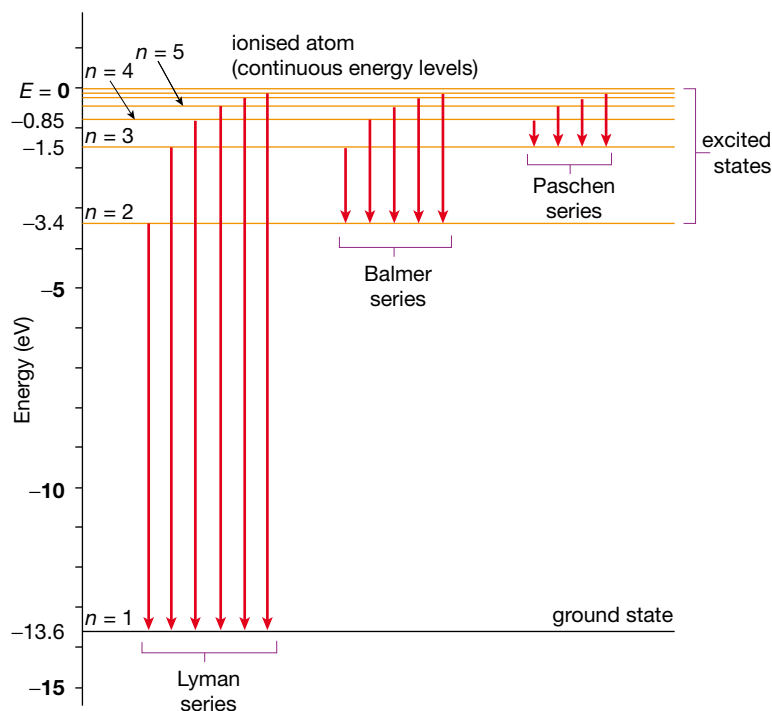
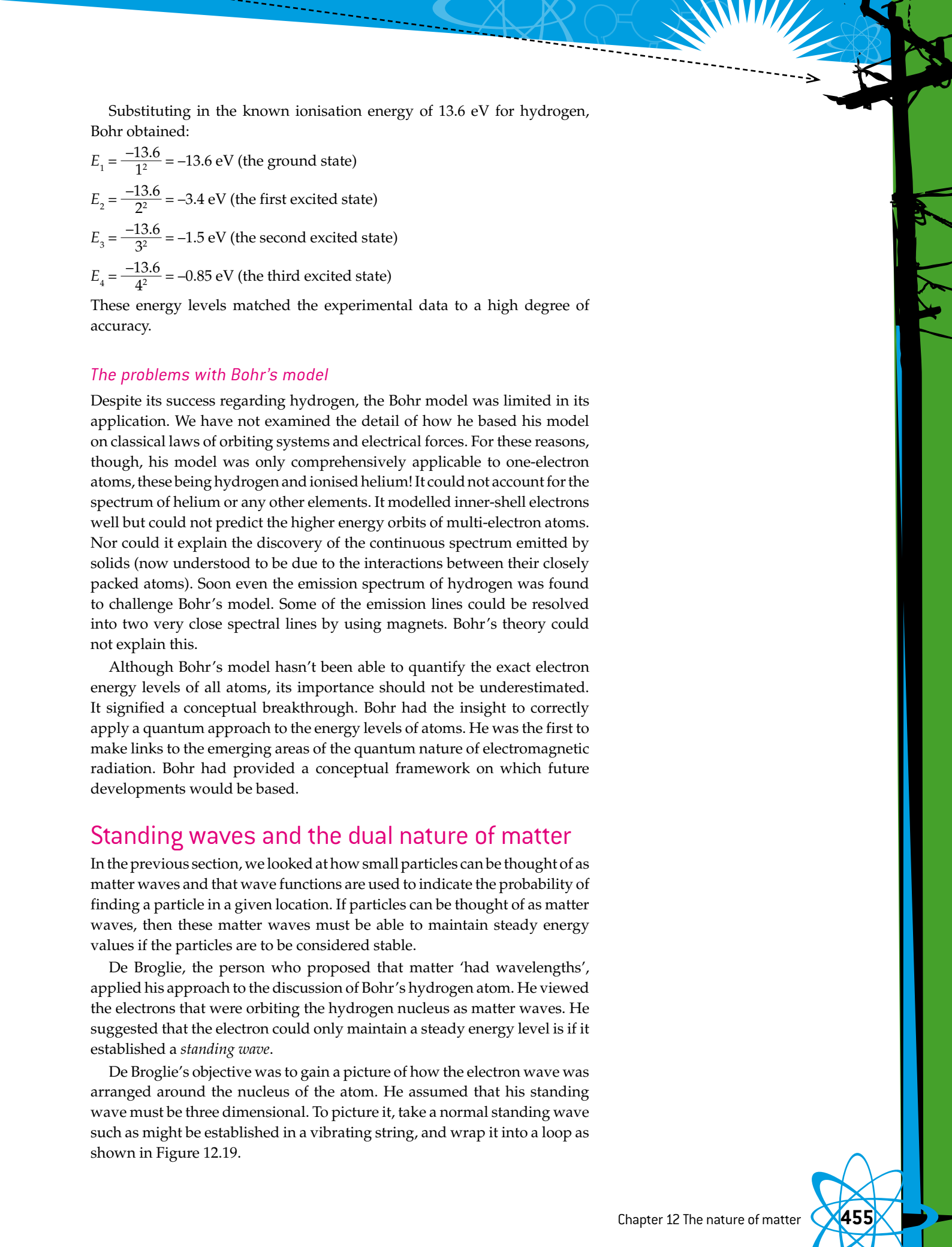


Figure 12.18 This is an energy level diagram for hydrogen. The ground state is bound by 13.6 eV and as one looks to higher energy levels, the energy levels are seen to crowd together.



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.

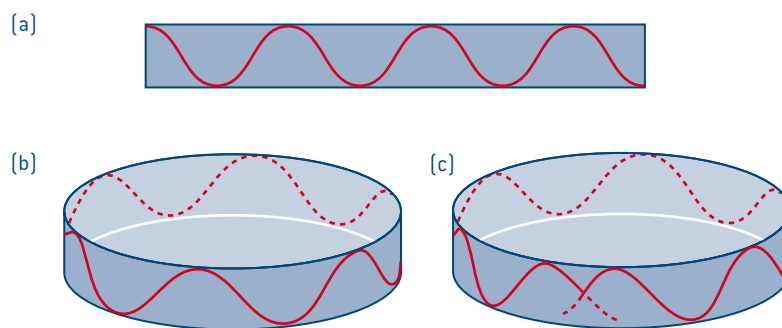
Standing waves and the dual nature of matter

In the previous section, we looked at how small particles can be thought of as matter waves and that wave functions are used to indicate the probability of finding a particle in a given location. If particles can be thought of as matter waves, then these matter waves must be able to maintain steady energy values if the particles are to be considered stable.

De Broglie, the person who proposed that matter 'had wavelengths', 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 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 12.19.

Figure 12.19 (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 could be equal to 1, 2, 3, ... De Broglie stated that $n\lambda = 2\pi r$ and substituting in his own matter wavelength expression, $\lambda = h/mv$, gives $nh/mv = 2\pi r$.

This was exactly the relationship that Bohr had assumed to exist (see Physics file, page 450) and that led to the expression that we examined for the energy levels of the electrons of the hydrogen atom, namely $E_n = -13.6/n^2$.

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 that followed him, such as Werner Heisenberg and Erwin Schrodinger. In our studies we have seen how early quantum ideas have explained the photoelectric effect, particle diffraction, 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



Figure 12.20 Although nuclear physics helped to bring World War II to a close, nuclear weapons testing continued.

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.

Physics in action

Standing waves

The expression 'standing wave' appears to be a contradiction, as surely a wave must involve movement of some type. In strings, this expression refers to a situation where a number of waves are interfering and result in the impression of a wave that is standing still. All physical objects have natural frequencies at which they can most readily vibrate. For strings, this depends on the length, tension and mass of the string; for air columns, on length and diameter. Whenever a musical instrument is played, it is forced by the musician to vibrate. If this forced frequency is equal to a natural frequency of the instrument, the amplitudes of the vibrations add together and resonance occurs. The resultant sound has a large amplitude; that is, it is quite loud.

The easiest way to gain an understanding of standing waves is to look at a string vibrating in its fundamental mode, as shown in Figure 12.22. When the string is vibrating in its most simple mode, the centre of the string is going through a maximum change in its displacement over a period of one cycle. At either end of the string there are fixed points. Hence, each of these points is at a displacement node. That is, they experience no change in their displacement value. Figure 12.22 shows a number of modes in which a string can vibrate, all related to the fundamental mode.

The fundamental standing wave is established so that the length of the string is equal to half a wavelength. Other modes of vibration are possible but



Figure 12.21 A close-up of a standing wave in a plucked guitar string.

all involve the wave envelope fitting exactly into the length of the string. The relationships between wavelength, λ , and string length, L , that is formed in each of the modes of vibration can be written:

$$L = \frac{1}{2}\lambda$$

$$L = 1\lambda$$

$$L = 1\frac{1}{2}\lambda$$

$$L = 2\lambda \text{ etc.}$$

It is clear from Figure 12.22 that any other modes of vibration cannot possibly exist, as a node must be present at each end of the string. The existing standing waves have a specific relationship to one another that can be expressed as:

$$\lambda_n = \frac{2L}{n}$$

first mode
(fundamental frequency)



$$\lambda_1 = 2L$$

second mode
(first overtone)



$$\lambda_2 = L$$

third mode
(second overtone)



$$\lambda_3 = \frac{2L}{3}$$

fourth mode
(third overtone)



$$\lambda_4 = \frac{L}{2}$$

Figure 12.22 A string is able to vibrate at a number of natural frequencies called harmonics. Each harmonic is a whole-number multiple of the fundamental frequency.



12.3 summary

Bohr, de Broglie and standing waves

- 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.



12.3 questions

Bohr, de Broglie and standing waves

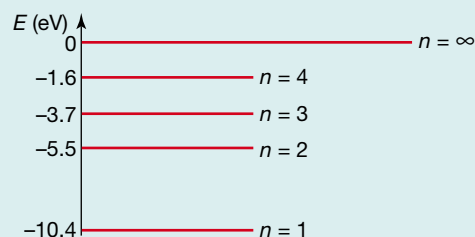
- Of the spectral lines in the Balmer series, which fall results in light of the shortest wavelength?
- Calculate the wavelength of the first three (longest wavelength) lines of the Lyman series.
- The visible spectrum extends from approximately 700 to 400 nm. Are all the Balmer lines in the hydrogen spectrum visible?
- What is the longest wavelength of EMR that can be absorbed by a hydrogen atom when an electron makes a transition from the ground state?
- What do de Broglie's matter wave concept and a bowed violin string have in common?
- Show how de Broglie's matter wave concept was important because it provided a model which vindicated some of Bohr's assumptions about atomic structure.
- Use Bohr's equation for the energy levels of hydrogen to calculate the 5th, 10th, 15th and 20th excitation levels for hydrogen.
 - What do you notice about the differences between the energy levels as n increases?



chapter review

The following information applies to questions 1 and 2.

The diagram shows the energy levels for atomic mercury.



- 1 An electron beam with energy of 8.8 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

- 2 List all the possible photon frequencies that could result when a mercury atom returns to the ground state from the second excited state.

- 3 The two-slit interference experiment can be duplicated by directing electrons through a copper foil. The resulting interference pattern is displayed using an electron detection screen.

- a What would be detected at a position of zero path difference?
b What would be detected at a point where the path difference is half a wavelength?
c Explain how electrons can exhibit this type of effect.

The following information applies to questions 4 and 5.

A 500 W lamp directs a beam of yellow light, of wavelength 580 nm, onto a perfect reflecting surface of area 4.0 cm^2 .

- 4 An electron beam is directed onto the reflecting surface. In order for the electron beam to exert the same pressure on the surface as the light beam, which of the following must be true?

A Each electron must have the same energy as a photon of yellow light.
B Each electron must have the same momentum as a photon of yellow light.
C The electrons must be travelling at the speed of light.

- 5 If the electrons in the beam each have the same energy as a photon of yellow light:

- a determine the momentum of each electron
b calculate the de Broglie wavelength of each electron
c determine the number of electrons that are incident on the surface each second in order to produce a pressure of 0.553 N m^{-2} .

- 6 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?

- 7 Describe the main features of the nuclear model of the atom as described by Rutherford.

- 8 Show how Bohr's model of the atom explained why:

- a the hydrogen atom was only capable of absorbing a small number of different frequencies of light
b hydrogen atoms have an *ionisation energy* of 13.6 eV.

- 9 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?

- 10 Give a description of how the standing wave model is considered to support for Bohr's theories about the atom.

exam-style questions Interactions of light and matter

For the following questions use:

$$e = 1.60 \times 10^{-19} \text{ C}$$

$$h = 6.63 \times 10^{-34} \text{ J s}$$

$$m_e = 9.11 \times 10^{-31} \text{ kg}$$

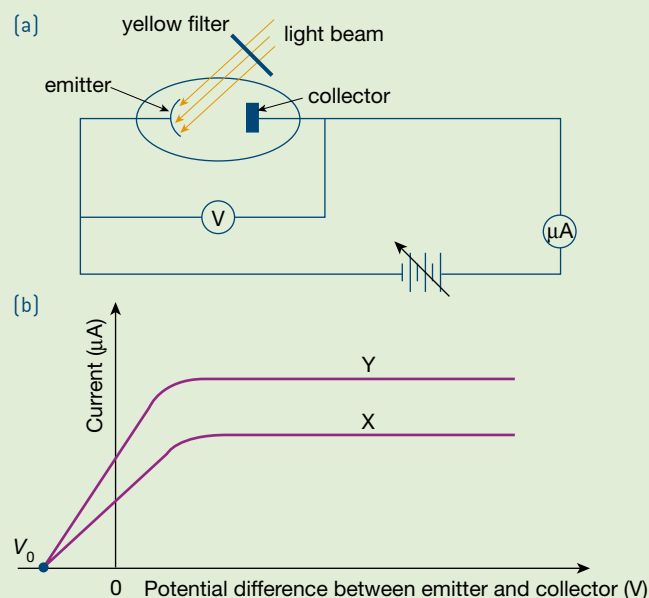
$$c = 3.00 \times 10^8 \text{ m s}^{-1}$$

$$m_p = 1.67 \times 10^{-27} \text{ kg}$$

- A 60 W lamp emits radiation of wavelength $3.0 \times 10^{-6} \text{ m}$.
 - How many photons per second is the lamp emitting?
 - What type of light is the lamp emitting?
- A beam of blue light, $7.0 \times 10^{14} \text{ Hz}$, is incident normally on a perfect reflecting surface. The beam power is 100 W.
 - Calculate the number of photons that are incident on the reflecting surface each second.
 - What is the momentum of each photon in the beam?
- A beam of yellow light, $5.2 \times 10^{14} \text{ Hz}$, is incident on a target which absorbs all the radiation falling on it. The beam delivers 1000 J of energy each second.
 - How many photons are striking the target each second?
 - What is the power of the beam?

The following information applies to questions 4–7.

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).



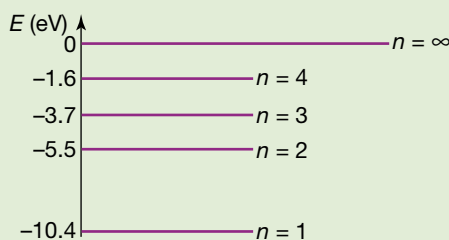
- Which of the following changes would result in an increase in the size of V_{out} ?
 - Replacing the yellow filter with a red filter
 - Replacing the yellow filter with a blue filter
 - Increasing the intensity of the yellow light

- 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?
 - The threshold frequency of the emitter increases to a value higher than the frequency of yellow light.
 - The work function of the emitter is increased to a value higher than the energy of a photon of yellow light.
 - The emitted photoelectrons do not have enough kinetic energy to reach the collector.
- Which of the following descriptions of the graphs X and Y are correct?
 - Both graphs are produced by yellow light of different intensities.
 - Graph X is produced by yellow light while graph Y is produced by blue light.
 - Each graph is produced by light of a different colour and different intensity.
- 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?
- Describe three experimental results associated with the photoelectric effect that cannot be explained by wave model theory of light.

The following information applies to questions 9 and 10.

In a double-slit interference experiment, an electron beam travels through two narrow slits, 20 mm apart, in a piece of copper foil. The resulting pattern is detected photographically at a distance of 2.0 m. The speed of the electrons is 0.1% of the speed of light.

- Calculate the de Broglie wavelength of the electrons used in the experiment.
 - What do you expect to see on the photographic plate?
 - Given that electrons are particles, how do you interpret the behaviour of the electrons in this experiment?
- If the experiment were to be repeated using neutrons, at what speed would a neutron need to travel to have the same de Broglie wavelength as the electrons in Question 9?
- 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 $n = 3$

12 An electron is accelerated across a potential difference of 65 V.

- a What kinetic energy will it gain?
- b What speed will it reach?
- c What is the de Broglie wavelength of the electron?

13 How would Niels Bohr explain the observation that for the hydrogen atom, when the frequency of incident light was below a certain value, the light would simply pass straight through the gas without any absorption occurring?

14 A muon is a particle with the same charge as an electron but its mass is 207 times greater. Consider an imaginary Bohr atom of a single muon orbiting a single proton, with energy levels at -1810 eV ($n = 1$), -703 eV ($n = 2$) and -312 eV ($n = 3$).

- a What is the ionisation energy of this atom?
- b What is the lowest frequency of light that would be able to ionise this atom?
- c What is the lowest frequency of light that could excite this atom?
- d Would incident light of energy 2000 eV be absorbed by this atom?
- e If a muon was travelling at 2% of the speed of light, calculate its de Broglie wavelength.

15 How would de Broglie explain the light and dark rings produced when a beam of electrons is fired through a sodium chloride crystal?

16 Describe how the wave—particle duality of electrons can be used to explain the quantised energy levels of the atoms.

17 Which one or more of the following phenomena can be modelled by a pure wave model of light?

- A the photoelectric effect
- B refraction
- C double-source interference of light
- D reflection
- E diffraction
- F the Compton effect

18 Define the electronvolt.

19 Why are all of the frequencies of light above the ionisation energy value for hydrogen continuously absorbed?

20 How do our models of waves and particles of light parallel the ideas related to electrons and matter waves?

21 Referring to Young's apparatus for his double-slit experiment, what is the effect on the observed diffraction pattern of:

- a halving the separation of the slits?
- b doubling the distance between the slits and the screen?
- c halving the frequency of the light used?
- d using white light?
- e covering one slit?

22 In a typical electron microscope, electrons are accelerated through a voltage of 10 kV.

- a Calculate the wavelength of an electron that has been accelerated through this potential difference.
- b Explain why an electron microscope is able to achieve a much greater degree of resolution than an optical microscope.

23 Would a 'proton microscope' be able to achieve a higher degree of resolution than an electron microscope with the same accelerating potential difference of 10 kV?

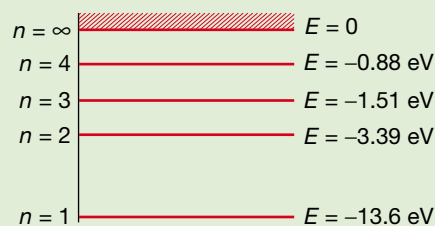
24 For an electron and a proton to have the same wavelength:

- A the electron must have the same energy as the proton
- B the electron must have the same speed as the proton
- C the electron must have the same momentum as the proton
- D it is impossible for an electron and a proton to have the same wavelength.

25 An X-ray photon of energy 18.0 keV collides with a stationary electron. The energy of the scattered photon is 16.0 keV.

- a Determine the kinetic energy of the electron after the collision.
- b What is the speed of the scattered photon?

26 Consider the energy level diagram for the hydrogen atom shown below. A photon of energy 14.0 eV collided with a hydrogen atom in the ground state.



- a Explain why this collision will eject an electron from the atom.
- b Calculate the energy of the ejected electron in electronvolts and in joules.
- c What is the momentum of the ejected electron?
- d Determine the wavelength of the ejected electron.

27 A hydrogen atom in the ground state collides with a 10.0 eV photon. Describe the result of such a collision, using the diagram in Question 26.

28 Consider the energy level diagram for the hydrogen atom shown in the diagram for Question 26. A collision between a 12.0 eV photon and a hydrogen atom in the ground state will:

- A result in the hydrogen atom being excited to the first excited state
- B result in the hydrogen atom being excited to the second excited state
- C ionise the hydrogen atom
- D leave the hydrogen atom in the ground state.

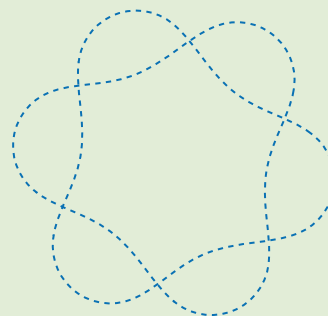
The following information applies to questions 29–32.

The energy levels for atomic mercury are shown below.

$n = \infty$	$E = 0$
$n = 4$	$E = -1.60 \text{ eV}$
$n = 3$	$E = -3.70 \text{ eV}$
$n = 2$	$E = -5.50 \text{ eV}$
$n = 1$	$E = -10.40 \text{ eV}$

- 29 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?
- 30 An electron beam of energy 7.0 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?
- 31 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?
- 32 Electrons of energy 4.0 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 $V = 6.2 \text{ V}$?

33 The diagram below represents the 'standing wave state' of an electron in an atom of hydrogen. Which value of ' n ' would de Broglie allocate to this pattern?



- 34 A beam of electrons of energy 8.00 eV passes through some mercury vapour in the ground state. According to the diagram in Question 29, which one of the following photon energies could not be produced from this interaction?
- A 2.10 eV
 - B 4.90 eV
 - C 1.80 eV
 - D 6.7 eV
- 35 A beam of photons of energy 8.00 eV passes through some mercury vapour. Discuss the likelihood of a mercury atom colliding with an 8.00 eV photon when it is already in an excited state.
- 36 What is the momentum of a gamma ray with a wavelength of 3.0 pm?

The following information applies to questions 37–40.

Figure 12.4 on page 441 shows the diffraction scattering images that have been obtained by scattering X-rays and electrons off the same sample, which is made up of many tiny crystals with random orientation. The X-rays have a frequency of $8.3 \times 10^{18} \text{ Hz}$.

- 37 Provide an explanation for the fact that the electrons and the X-rays have produced the same diffraction pattern.
- 38 Determine the wavelength of the X-ray photons.
- 39 Determine the wavelength of the electrons.
- 40 Calculate the momentum of the electrons.

Unit

area of study 3

4

Detailed studies

detailed studies

Chapters 13–15 are the detailed studies for Unit 4. You will undertake one detailed study in each unit.

Chapter 13 Synchrotron and applications
Chapter 14 Photonics
Chapter 15 Sound

Synchrotron and applications

Melbourne became home to the most powerful synchrotron in the southern hemisphere in July 2007. The synchrotron stands on the old Clayton drive-in site, on the corner of Wellington and Blackburn roads, next to the main campus of Monash University. 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.

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 supermicroscopes 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.'



outcome

On completion of this chapter, you should be able to describe the basic design and operation of the Australian Synchrotron and the production, characteristics and interactions with targets of synchrotron radiation.



by the end of this chapter

you will have covered material from the study of the synchrotron and its applications including:

- 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
- analysis of data from experiments that involve interactions of synchrotron light with a sample, including X-ray (Bragg) diffraction, emission of photoelectrons and production of X-ray absorption spectra
- a description of types of X-ray scattering including elastic (Thomson) scattering and inelastic (Compton) scattering.

13.1 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 channeled off for use in experimental stations to analyse materials.

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 13.5 is called a Maltese cross tube. Inside this, electrons are accelerated from the hot cathode

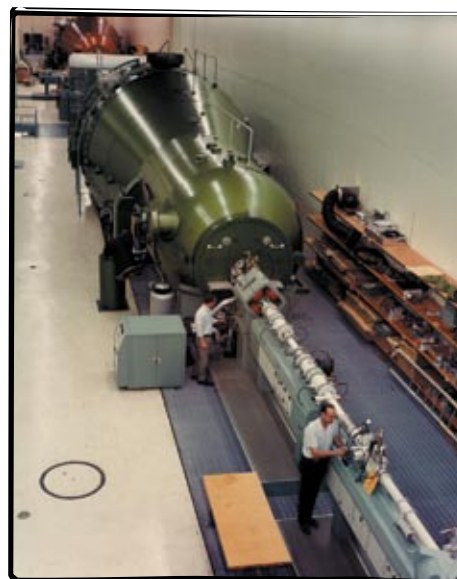


Figure 13.1 This tandem van de Graaff generator uses two generators to produce beams of charged particles that are accelerated by potential differences of up to 10 million volts.



Figure 13.2 Electrons are accelerated from the heated filament within an incandescent light bulb which produces photons of light.



Figure 13.3 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.

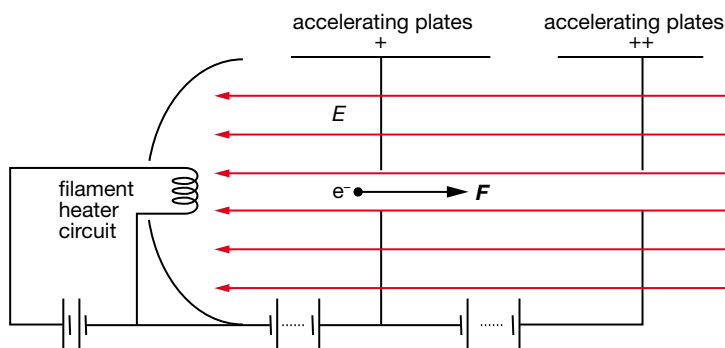


Figure 13.4 The set-up of a typical electron gun assembly. Electrons are released from a hot cathode and are accelerated across a potential difference through slits in a pair of positive charged plates. A magnetic field helps to centralise the path of the electrons.

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.

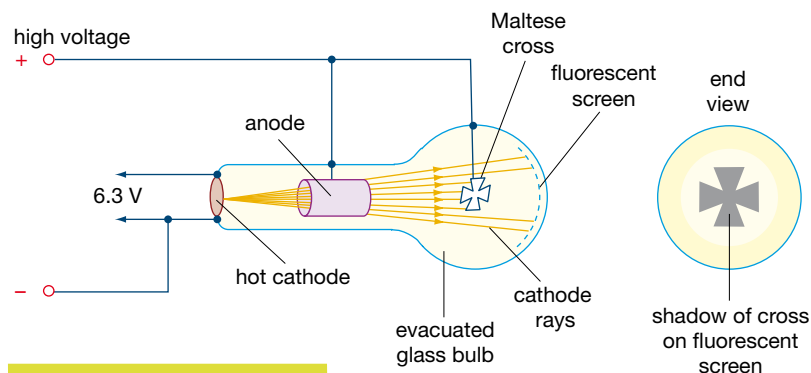


Figure 13.5 The Maltese cross tube.

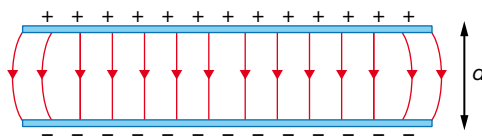


Figure 13.6 The electric field between a pair of oppositely charged plates. The direction of the field is defined as being the direction in which a positive charge would move in this region.

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:

$$E = \frac{F}{q}$$

where F is the force (N) experienced by a charged particle due an electric field 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:

$$E = \frac{V}{d}$$

where d is the separation of the plates (m) and V 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:

$$F = \frac{qV}{d}$$

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:

$$W = qV$$

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:

$$\begin{aligned}\Delta E_k &= W \\ &= qV\end{aligned}$$

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 9 and 10, you will remember that because an electric current is itself a stream of moving charges, we can state that the magnitude of the force, F , on a charge, q , moving with velocity v perpendicular to a magnetic field of strength, B , is given by:

$$F = qvB$$

So, in the case of the magnetic force on an electron moving within the magnetic field of a particle accelerator:

$$F = evB$$

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 $F = evB\sin\theta$, where θ is the angle between the direction of velocity and the magnetic field.

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 .

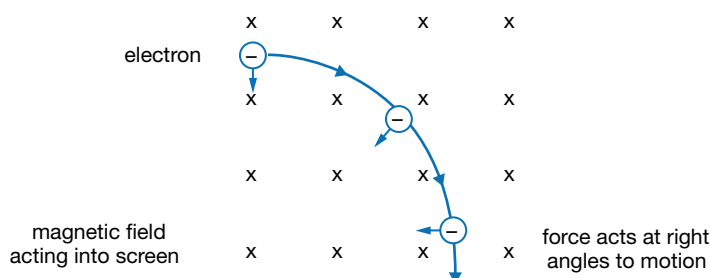


Figure 13.7 An electron fired horizontally to the right into a magnetic field that acts into the page will initially experience a force vertically downwards as predicted by the right-hand palm rule. It will follow a curved path of radius r . Note that the direction as defined for the current in the right-hand palm rule must be reversed to provide the direction that an electron will follow.

Physics file

You will recall from Figure 12.7, page 443, that the British scientist J.J. Thomson used an apparatus such as this arrangement 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 electrons, Thomson calculated charge-to-mass ratio to be $1.759 \times 10^{11} \text{ C kg}^{-1}$. 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.

In this case we can say that the net force acting on the charge is:

$$F = ma$$

This is equivalent to the magnetic force on the charge, so that:

$$ma = evB$$

The acceleration in this situation is centripetal and has magnitude:

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

Substituting this value into the previous equation:

$$\frac{mv^2}{r} = evB$$

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:

$$r = \frac{mv}{eB}$$

As the momentum of the charged particle is equivalent to $m\mathbf{v}$, we can also state this radius as:

$$r = \frac{p}{eB}, \text{ where } p = \text{the momentum of the electron (kg m s}^{-1}\text{)}.$$

Worked example 13.1A

An electron gun releases electrons from its cathode, which are accelerated across a potential difference of 32 kV, a distance of 30 cm between a pair of charged parallel plates. [Assume that the mass of an electron is $9.1 \times 10^{-31} \text{ kg}$, the charge on an electron is $-1.6 \times 10^{-19} \text{ C}$ and ignore any effects of relativity in your calculations.]

- Calculate the strength of the electric field acting on the electron beam.
- Calculate the magnitude of the velocity of the electrons as they exit the electron gun assembly.
- The electrons then travel through a uniform magnetic field perpendicular to their motion. Given that this field is of strength 0.2 T, calculate the expected radius of the path of the electron beam.

Solution

$$\begin{aligned} \text{a } E &= \frac{V}{d} \\ &= \frac{32 \times 10^3}{0.3} \\ &= 1.1 \times 10^5 \text{ V m}^{-1} \end{aligned}$$

$$\begin{aligned} \text{b } \frac{1}{2}mv^2 &= eV \\ v^2 &= \frac{2eV}{m} \\ \Rightarrow v &= \sqrt{\frac{2 \times 1.6 \times 10^{-19} \times 32 \times 10^3}{9.1 \times 10^{-31}}} \end{aligned}$$

$$v \text{ is approximately } 1.1 \times 10^8 \text{ m s}^{-1}$$

$$\begin{aligned} \text{c } r &= \frac{mv}{eB} \\ &= \frac{9.1 \times 10^{-31} \times 1.1 \times 10^8}{1.6 \times 10^{-19} \times 0.2} \\ &= 3.1 \times 10^{-3} \text{ m} \end{aligned}$$

This means the electrons will follow a path of radius 3.1 mm. In actual fact, this extremely small radius is not realistic, due to the effects of relativity, which will be considered shortly.

Linear accelerators

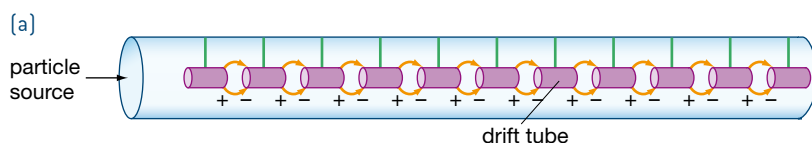


Figure 13.8 [a] Charged particles entering the linac from the left are accelerated towards the drift tubes by an electric field. They travel through the drift tubes at constant velocity, before being accelerated again between the gaps of each tube. [b] This is the standing wave linear accelerator found at the Berkeley Laboratory. Can you see the drift tubes?

Cathode ray tubes are useful particle accelerators but are limited to using voltages over 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 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.

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 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.

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:

$$r = \frac{mv}{qB}$$

Their speed is then $v = \frac{rqB}{m}$ and the time taken for one orbit

of the cyclotron is: $t = \frac{d}{v}$ where d = path distance of one revolution

$$\begin{aligned} &= \frac{2\pi r}{v} \\ &= \frac{2\pi mv}{qBv} \text{ (substituting the above expression for path radius, } r\text{)} \\ &= \frac{2\pi m}{qB} \end{aligned}$$

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.

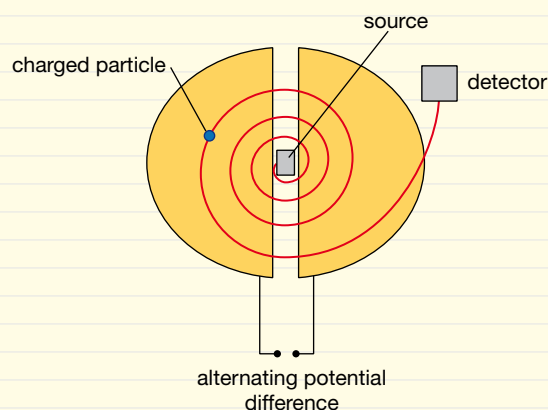


Figure 13.9 The cyclotron operates in some ways like a spiral linac—particles are accelerated from a source, through semicircular chambers called dees, until they gain sufficient energy to exit.

Physics 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 13.1 is a sample of radioisotopes produced in cyclotrons and their uses.

Table 13.1 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

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 (1TeV = 1000 GeV) each,



Figure 13.10 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 13.11 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 World Wide Web itself originated as a result of physicists at CERN being interested in developing a method of sharing information across long distances.

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.

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.



13.1 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 $F = qE$, where E is the electric field strength. This can be rewritten as $F = qV/d$ as the electric field strength is equivalent to V/d , where V is the potential difference between plates and d is the plate separation.

- The work done in an electron of charge q accelerating across this potential difference V is:

$$W = qV$$

- The increase in the kinetic energy of an electron of mass m with final velocity v is:

$$\Delta E_k = qV$$

- For the case of an electron moving at right angles to a magnetic field, the force it experiences is:

$$F = qvB$$

If the magnetic force is the net force acting on the electron, then it will move in a circular path of radius r :

$$r = \frac{mv}{qB}$$

- 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.



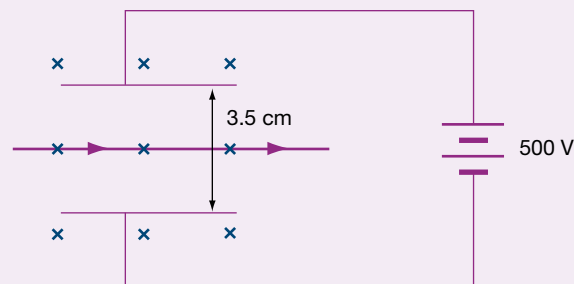
13.1 questions

Particle accelerators

For each of the following questions, assume that the mass of an electron is 9.1×10^{-31} kg and its charge is 1.6×10^{-19} C.

- Electrons in a cathode ray tube are released from a:
 - hot anode
 - hot cathode
 - cool cathode
 - cool anode.
- Sir William Crookes developed a number of tubes to study cathode rays.
 - Describe the basic set-up of these cathode ray tubes.
 - Describe how the electrons are accelerated through the tubes.
- Describe the operation of the standing-wave linac. Why is this type of linac not employed in the Australian Synchrotron?
- 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?
- An electron gun assembly emits electrons with energies of 10 keV. Ignore the effect of relativity in answering the following questions.
 - Calculate the magnitude of the predicted exit velocity of the electrons:
 - Upon exiting the electron gun assembly, the electrons now enter a uniform magnetic field of 1.5 T acting perpendicular to their motion. Calculate the predicted radius of the electron beam.
- This diagram represents an electron being fired at right angles towards a uniform magnetic field acting out of the page.

- Copy the diagram and mark on it the continued path you expect the electron will follow.
 - 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 electric field strength between the plates
 - the speed of the electrons, given that the magnetic field is of flux density 1.5×10^{-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 2.5 kV.
- 9 An electron with speed of 7.6×10^6 m s⁻¹ travels through a uniform magnetic field and follows a circular path of diameter 9.2×10^{-2} m. Calculate the magnetic flux density of the field through which the electron travels.
- 10
- Calculate the force exerted on an electron travelling at speed of 7.0×10^6 m s⁻¹ at right angles to a uniform magnetic field of strength 8.6×10^{-3} T.
 - Given that this force directs the electron in a circular path, calculate the radius of its orbit.

13.2 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.

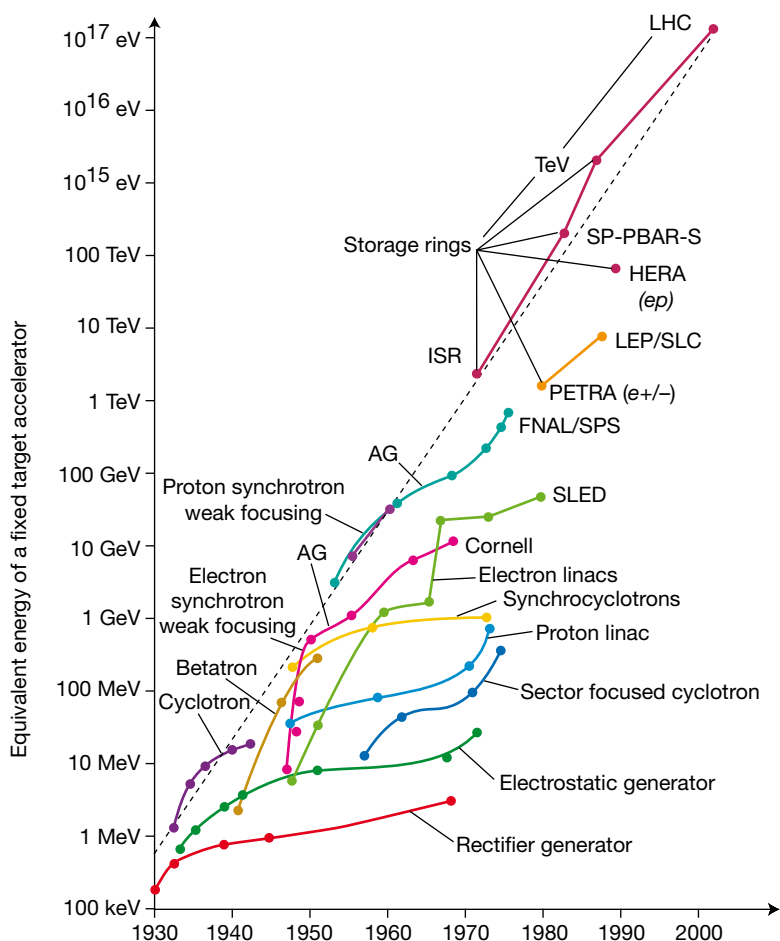


Figure 13.12 This graph demonstrates the rapid rise in capabilities of particle accelerators dating from the 1930s to the present day.

How do they work?

Figure 13.13 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.

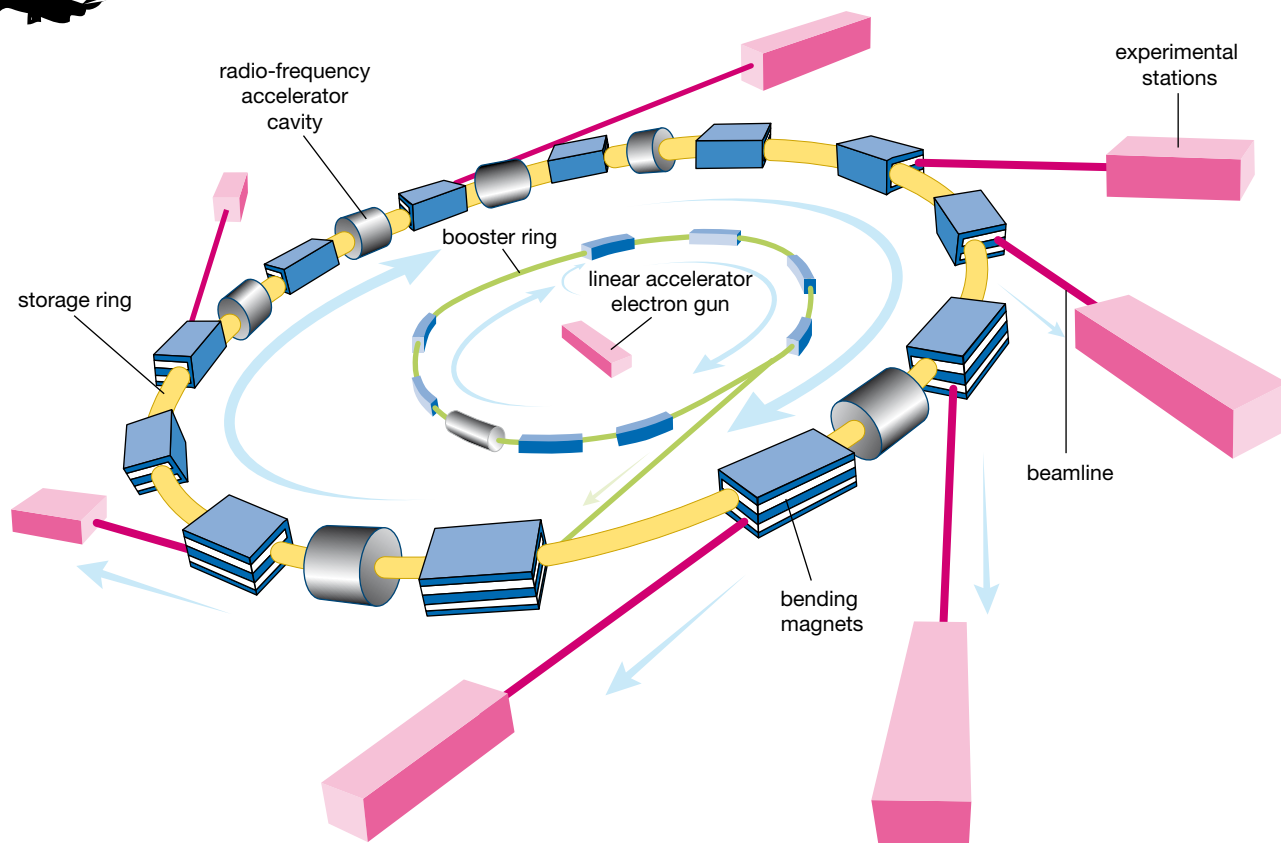


Figure 13.13 The Australian Synchrotron is almost the size of the MCG oval. This scale of some 67 m in diameter is necessary to contain the electrons which are travelling at almost the speed of light as they zoom around the storage ring. So, how does it actually work?

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

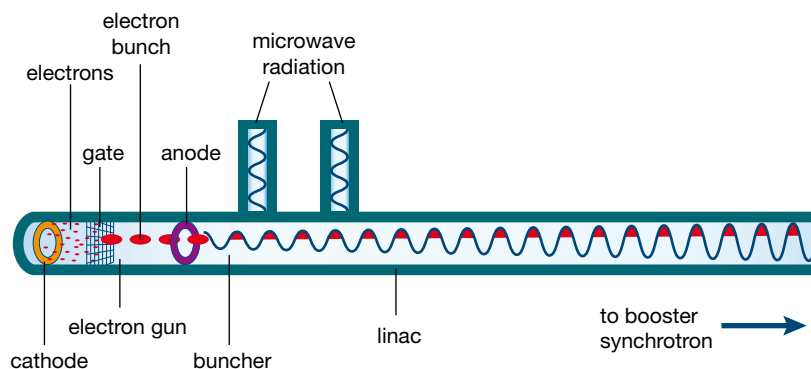


Figure 13.14 A basic travelling-wave linac design, showing the electrons being emitted from the electron gun, accelerating towards an anode and then being further accelerated in bunches by RF radiation.

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 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 ($c = 3 \times 10^8$ 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 as we give the electrons more energy. 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.

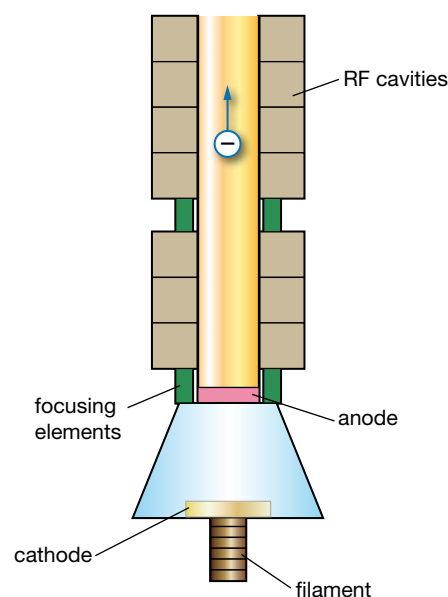


Figure 13.15 In the Australian Synchrotron, electrons boil off the heated filament of the electron gun assembly. The gate in this diagram is a grid that is also a cathode. A bias voltage prevents the electrons from travelling towards the anode. A 500 MHz RF voltage is applied every 2 ns to overcome the bias voltage, allowing electrons to travel in pulses towards the anode.



Figure 13.16 Looking like a mass of metal components, this photograph shows part of the RF system of the Australian Synchrotron.



Figure 13.17 The booster to storage ring transfer line in the Australian Synchrotron.

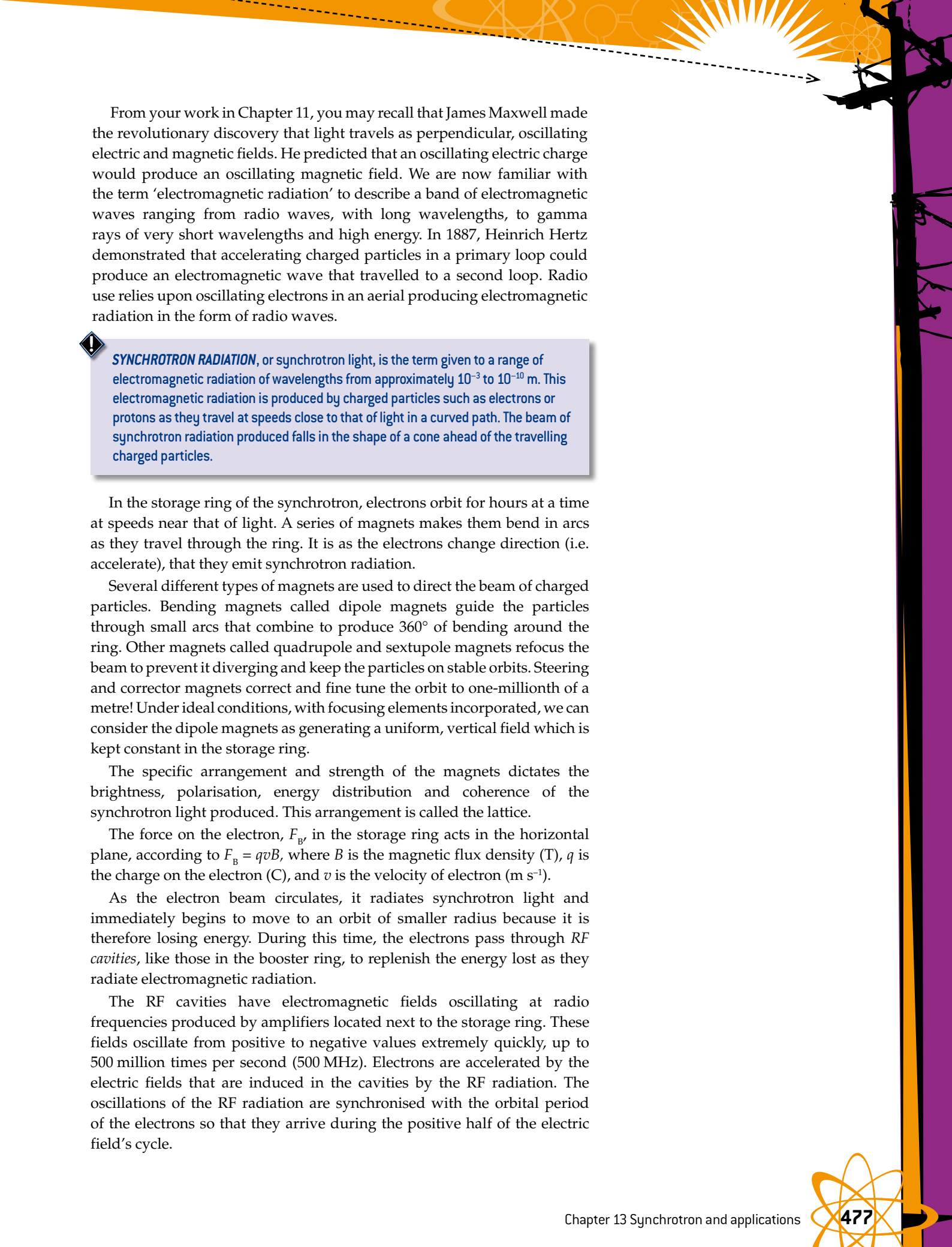
Einstein found that the effective mass of an object increases very sharply as its speed approaches c . Even at 10% of c , the mass has only increased by about 0.5%, at 87% it has doubled, and at 99% of c the mass has increased by about seven times. After 99% of c , the mass increases very rapidly. In 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.

Recall also that the radius of the path of the electrons is given by the expression $r = \frac{mv}{eB}$. 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) 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 $r = \frac{mv}{eB}$, 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.



From your work in Chapter 11, 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, F_B , in the storage ring acts in the horizontal plane, according to $F_B = qvB$, where B is the magnetic flux density (T), q is the charge on the electron (C), and 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.

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.

This is where the term 'synchrotron' originates. This process ensures that the electrons stay at a constant energy and remain stored in the ring. 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

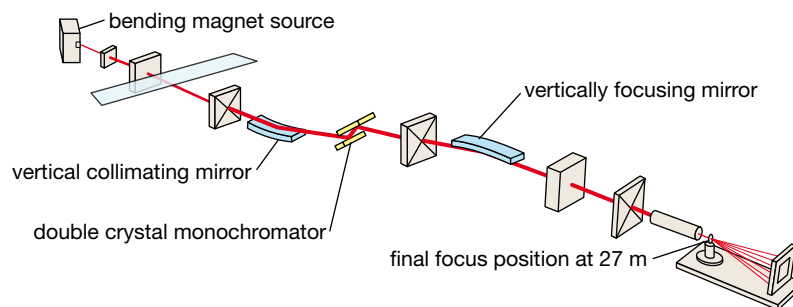
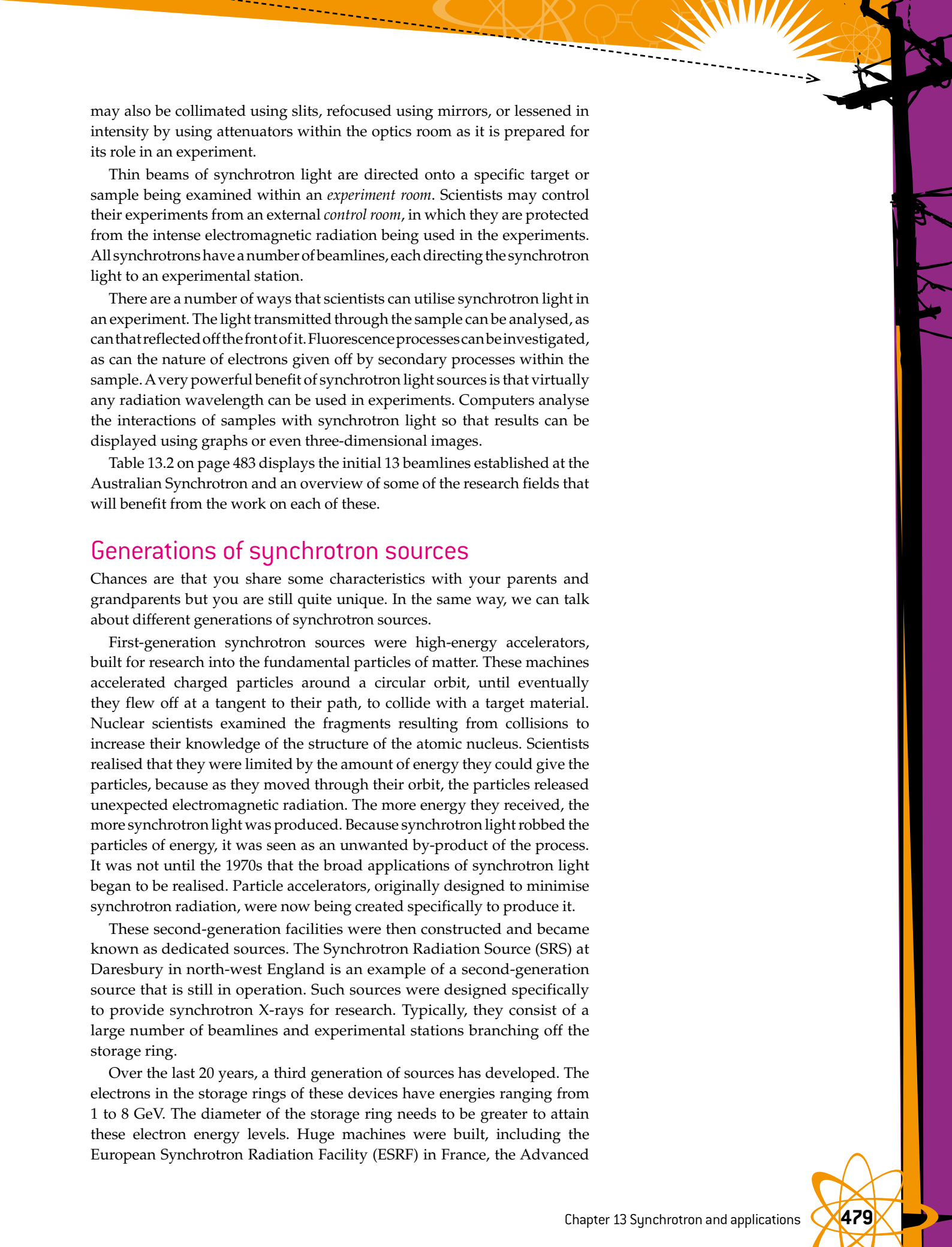


Figure 13.18 This diagram shows the arrangement of mirrors and crystal monochromator in the protein and microcrystal and small molecule X-ray diffraction beamline in the Australian Synchrotron.

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 13.2 on page 483 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

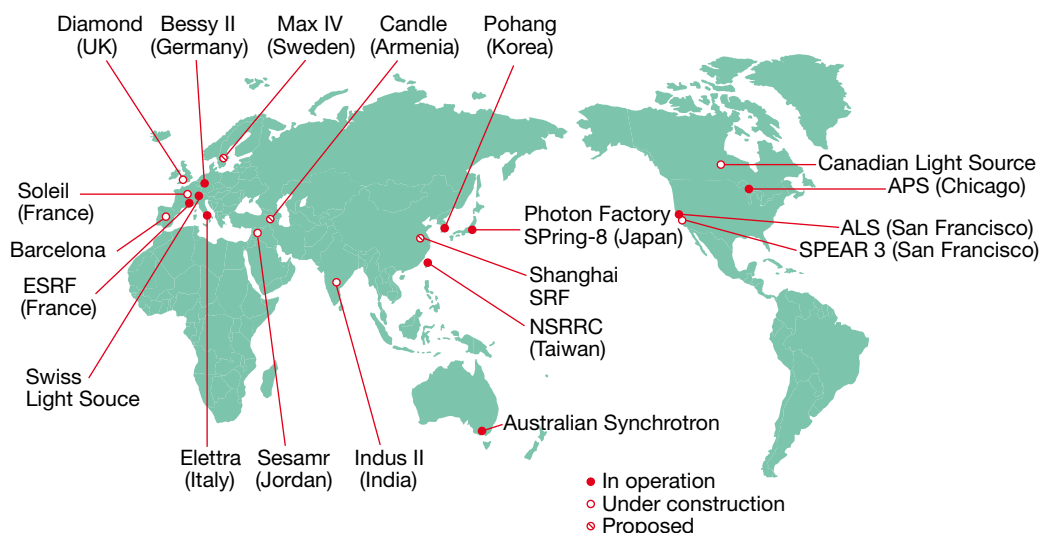


Figure 13.19 This map identifies third-generation synchrotron facilities with energies greater than 1.5 GeV currently proposed and in operation around the world.

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.

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.

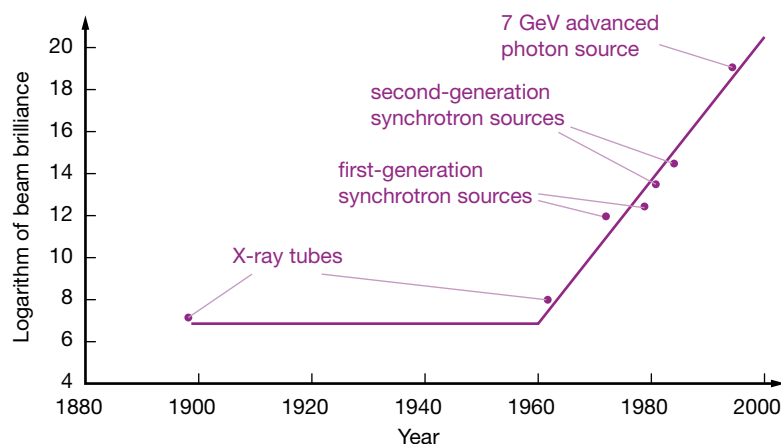


Figure 13.20 The brightness of subsequent generations of X-ray sources has greatly increased over the past 50 years.

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 13.21a.

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 13.21b.

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 13.21c. 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.

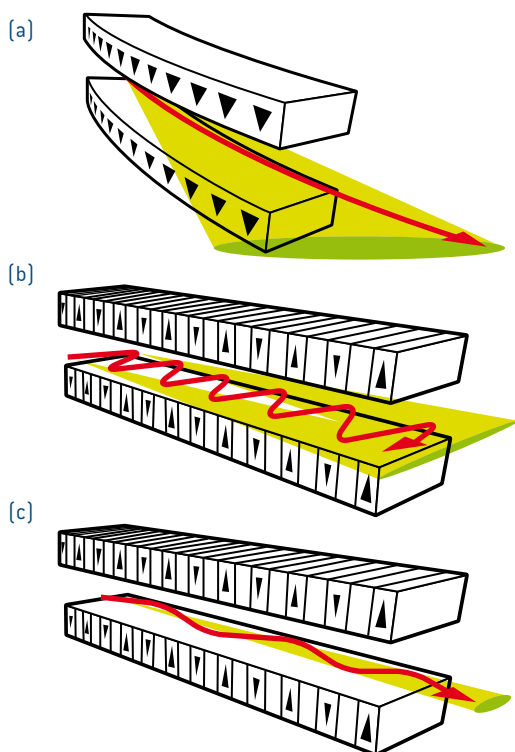


Figure 13.21 These diagrams show the bending of the path of the electrons and the resultant band of synchrotron light produced using (a) a dipole magnet, (b) a wiggler and (c) an undulator.

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 13.2, page 483).

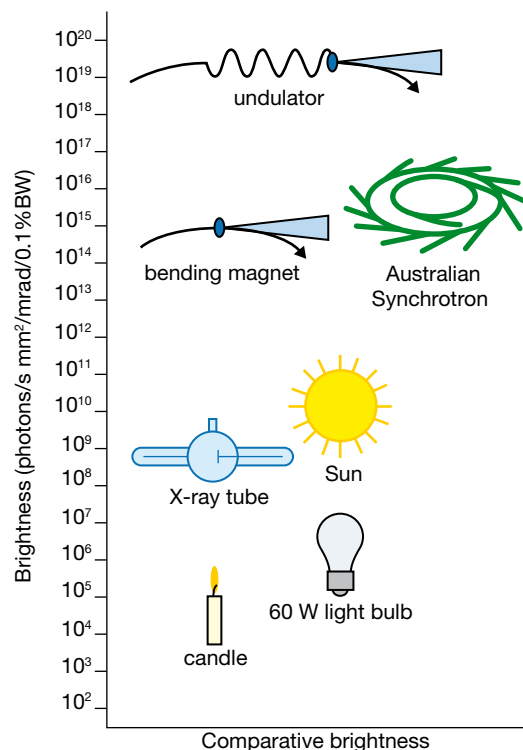


Figure 13.22 Compare the brightness of electromagnetic radiation produced by more traditional sources with respect to what is possible using insertion devices within a synchrotron facility.

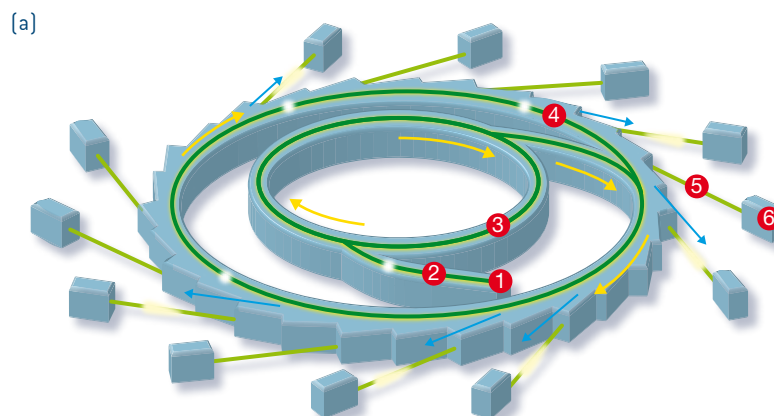


Figure 13.23 (a) The design of the Australian Synchrotron: (1) electron gun, (2) linac, (3) booster ring, (4) storage ring, (5) beamline, (6) experimental station. (b) View of the inside of the Australian Synchrotron, taken from the mezzanine.

Table 13.2 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 & optical spectroscopy	Microfocus spectroscopies	Imaging and medical therapy	General purpose microprobe	Circular dichroism	Lithography
Life sciences													
Biological research & drug design	•	•	•	•	•			•	•	•	•	•	
Biotechnology & bio-sensors	•	•			•	•		•	•				•
Biomedical & medical imaging								•	•	•			
Medical therapy										•			
Plants and crops	•		•		•		•			•		•	
Physical sciences													
Sustainable environment		•	•		•			•	•				
Forensics			•	•	•	•		•	•	•			
Advanced materials													
– functional polymers		•		•		•		•		•		•	
– ceramics			•	•	•	•		•	•	•	•		
– nanomaterials & composites		•	•	•		•	•	•	•	•	•		•
– metals & alloys			•			•	•		•	•	•		
– micro-electronic & magnetic materials					•	•	•	•		•			•
– biomaterials		•	•	•	•	•	•	•		•	•		
Engineering			•							•			•
Mineral exploration & beneficiation			•		•	•			•	•	•		
Earth sciences		•	•	•	•	•		•	•		•		
Oil and gas production and distribution		•		•	•			•	•	•	•		
Agricultural technology	•		•	•	•	•				•		•	•
Food technology		•	•	•				•				•	•
Chemical reactions & catalysts		•	•	•	•	•	•	•	•		•	•	•



13.2 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.



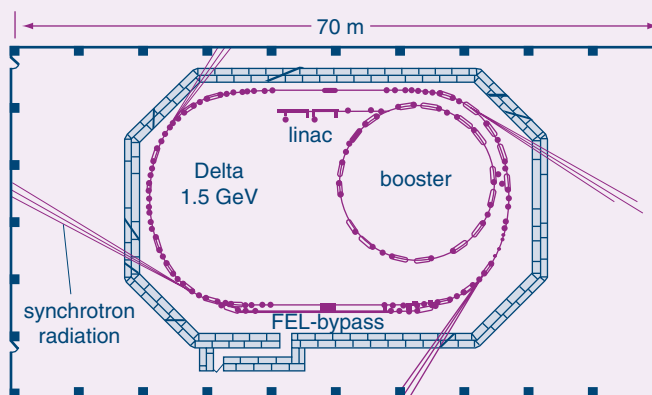
13.2 questions

Synchrotrons

- 1 a** Carefully describe the function of these major parts of the synchrotron:
 - i the linac
 - ii the circular booster ring
 - iii the storage ring
 - iv the 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 482) 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 13.2, page 483. 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 proposed Australian Synchrotron. List two similarities and two differences.



13.3 Synchrotron radiation

What is synchrotron light?

We have traced the development of increasingly powerful synchrotron light sources. But what is synchrotron light and what are its applications?

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
- tunable—required frequencies can be selected from synchrotron light
- pulsed—emitted in very short pulses of less than a nanosecond
- highly polarised—either linearly, circularly or elliptically polarised.

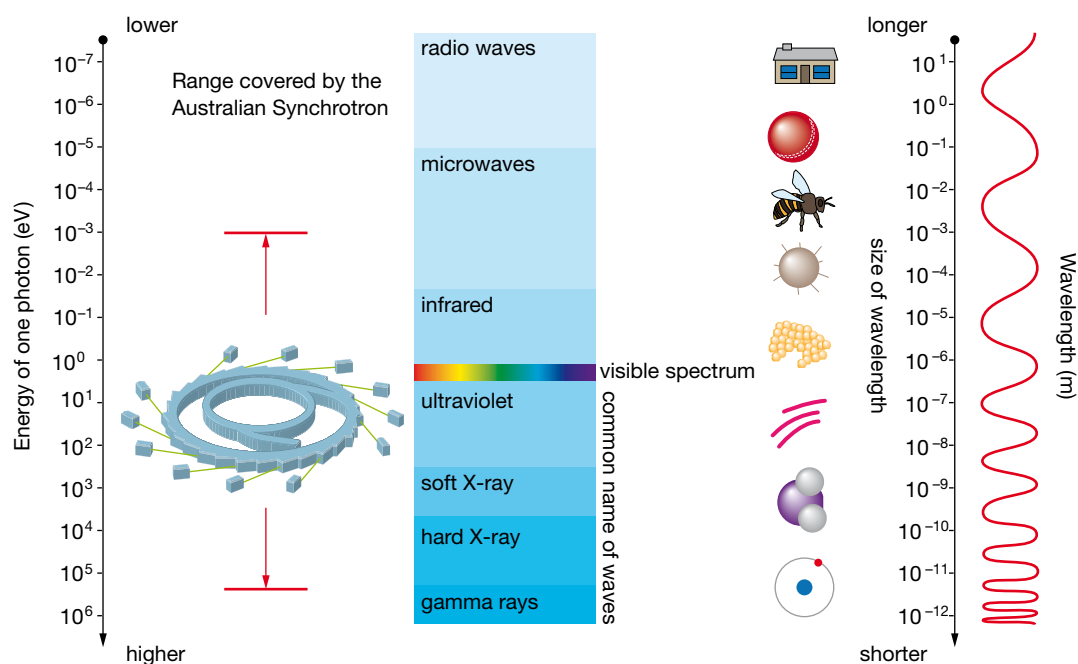


Figure 13.24 The range of wavelengths and photon energies produced as synchrotron radiation.

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 13.24. Such electromagnetic radiation has wavelengths corresponding to the dimensions of cells, viruses, proteins and atoms. From Figure 13.24 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.

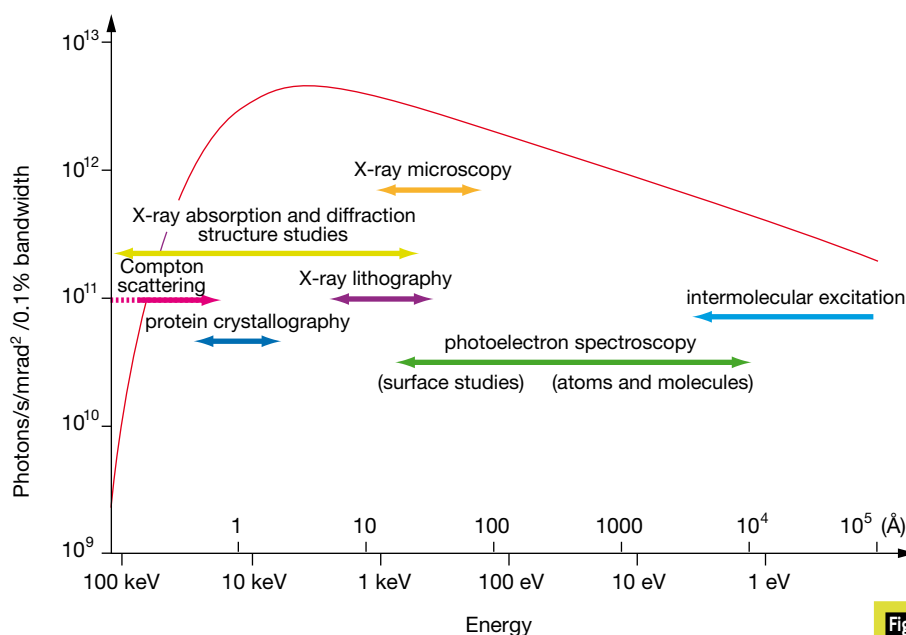


Figure 13.25 Almost all areas of science have benefited from the use of synchrotron radiation. The range of techniques possible using various bands of synchrotron radiation are highlighted in this figure. This graph illustrates the range of synchrotron radiation emitted from the Synchrotron Radiation 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, do 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 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

X-rays were so named because their properties were initially unknown. They were first discovered by Wilhelm Conrad Röntgen in 1895.

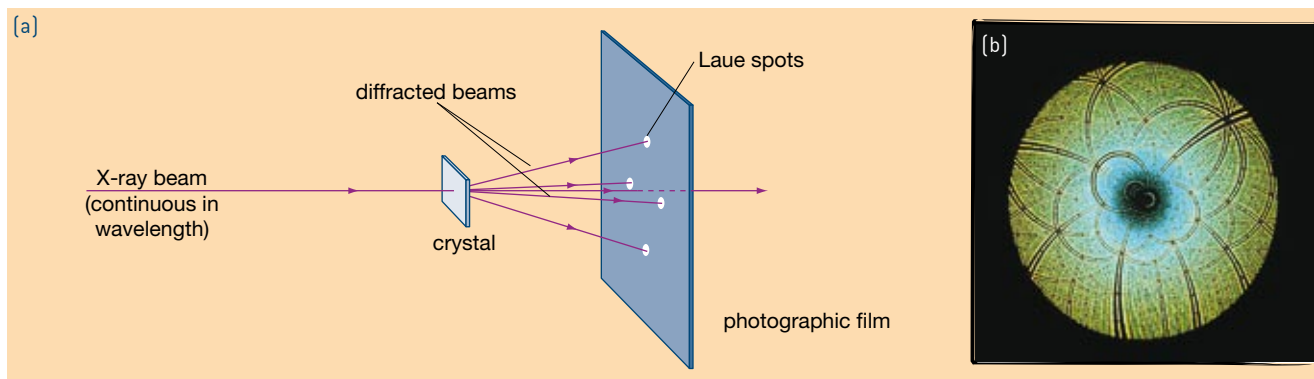


Figure 13.26 (a) The arrangement of atoms in the crystal can be determined from an analysis of the Laue spots produced in the diffraction pattern. (b) Laue X-ray diffraction pattern from a crystal of the enzyme rubisco [ribulose biphosphate carboxylase oxygenase] from a plant.

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 13.26a.

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 13.26b). 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 least of all 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 13.27 Say cheese! The toothy grin of a dental X-ray.

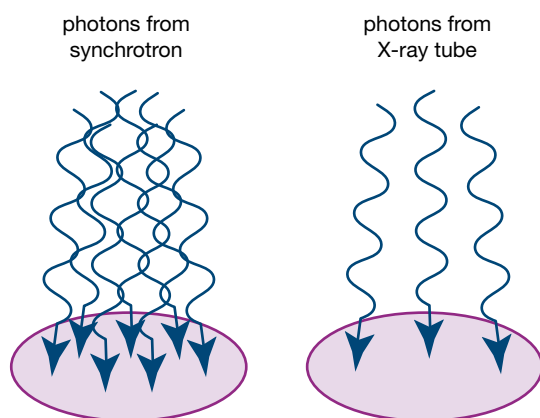


Figure 13.28 The X-rays delivered from a synchrotron source are far brighter than X-rays produced by an X-ray tube because a higher intensity is concentrated in a smaller area.

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 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 13.28. 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.

Lithography

We live in a world where everyday appliances are getting smaller. The ability to manufacture extremely small devices could lead to groundbreaking new applications in medicine, microsensors and micromachining. Lithography consists of printing a pattern onto another surface. Lithographic techniques have existed for centuries. The method of using X-rays to etch a pattern is much more recent. X-ray lithography is an exciting application of synchrotron radiation. Lithographic techniques can be used to create tiny mechanical parts or to produce moulds that may be filled with a range of materials, including metals, glass or ceramics, to make components for devices. This technique is capable of producing high-resolution images between $100\text{ }\mu\text{m}$ and 2 mm thick.

One application is the production of micro-electrical mechanical devices such as microsensors and actuators, for example, motion devices, pumps and filters. Synchrotron light is suitable for providing the X-rays used in the lithographic process.

X-rays of wavelength $0.01\text{--}1.0\text{ nm}$ are generally selected from the beam and are shone through a patterned mask onto a photoresist coating that is later developed to reveal the image. This process is shown in Figure 13.29. As the

radiation is exposed, it casts a shadow of the mask pattern onto the resist. X-rays interact with the resist coating.

This photosensitive layer produces images of resolution currently approaching $0.14\text{ }\mu\text{m}$. Electrons produced from the photoelectric effect induce a change so that exposed resist can be removed by chemical development, while unexposed resist is left behind. Imaging systems called X-ray aligners, or steppers, are used to produce a clearer image. They operate on a vertically held mechanical stage.

The collimation of the X-ray beam produced in synchrotron radiation makes it particularly suited to the technique of X-ray lithography. The shadow-blurring effect due to light incoherence is reduced. Because the beam is of high intensity, shorter exposure times are used, which prevents overheating of the mask arrangement. The suitability of synchrotron light to this process has resulted in the fabrication of much smaller, more compact storage rings specifically to perform X-ray lithography work. A process of deep-etch lithography, called LIGA, can be used to produce gears only a few thousandths of a millimetre in size. These components, along with microscopic turbines, can be used to build microelectronic devices in the revolutionary new field of micromachining.

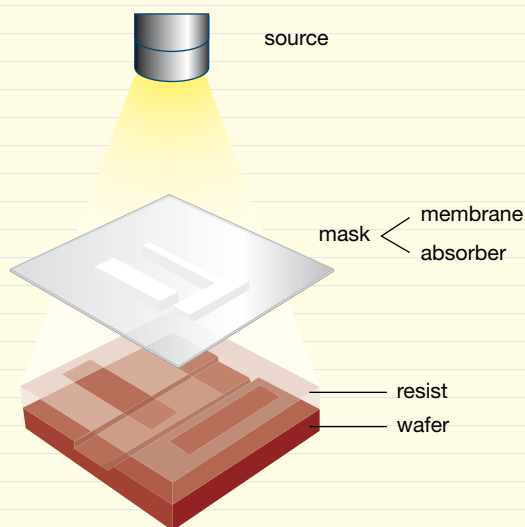


Figure 13.29 The basic set-up of an X-ray lithographic process. The mask used in the process consists of a transparent part called the carrier. This is a thin membrane, generally $1\text{--}2\text{ }\mu\text{m}$ thick, made from silicon carbide or silicon nitride. The pattern to be copied is structured in a thick layer of a heavy element that is attached to this membrane. This section is called the absorber, and is usually made up of gold or tungsten. X-rays are absorbed into the absorber and pass through the membrane in order to transmit the desired pattern to the wafer below.



Figure 13.30 Gears like this one can be produced by using deep-etch lithographic techniques. They are used to build micromachines.

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 for 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 13.31 Sir Lawrence Bragg (1890–1971)—known as the father of the science of X-ray crystallography.

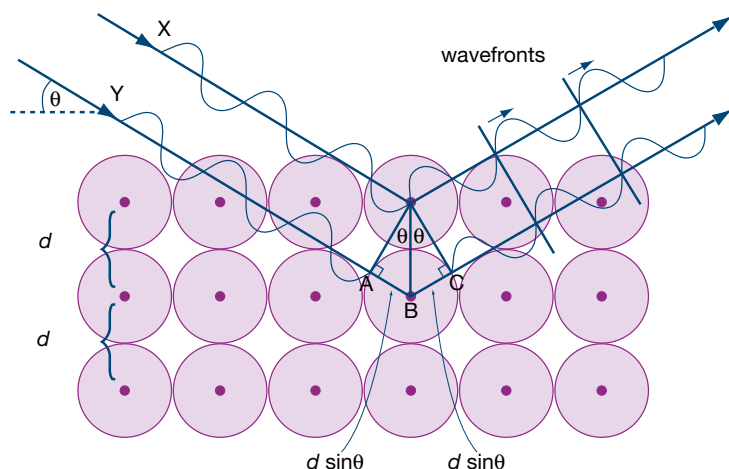


Figure 13.32 Wavefronts X and Y are scattered from layers of atoms of separation, d .

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 13.32. 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.



BRAGG'S LAW states that for constructive interference:

$$d\sin\theta + d\sin\theta = n\lambda$$

$$2d\sin\theta = n\lambda$$

where d = the distance between layers of atoms (m)

θ = the angle the X-ray beam makes with the crystal surface (called the glancing angle)

λ = the wavelength of incident X-ray photons (m)

n = the number of maxima occurring, 1, 2, 3, etc.

Reflected beams of X-rays of a fixed wavelength will then occur for the glancing angles given by this rule.

Increasing θ will allow corresponding higher order maxima for $n = 2, 3$, etc. Substituting each glancing angle into Bragg's law enables us to calculate d , the distance between layers of atoms, for a given crystal sample. In this way, scientists gain valuable information about the structure of the sample crystal.

Worked example 13.3A

A crystal of sodium chloride is bombarded with X-rays, and the first-order diffraction of the beam is observed at a glancing angle of 12.3° . Given that the X-rays had wavelengths of 120 pm, calculate the distance between adjacent planes of ions by using Bragg's equation.

Solution

For a first-order maximum, $n = 1$. So, Bragg's equation, $2d\sin\theta = n\lambda$, becomes $2d\sin\theta = \lambda$. Substituting values:

$$\begin{aligned} d &= \frac{\lambda}{2\sin\theta} \\ &= \frac{120 \times 10^{-12}}{2\sin 12.3^\circ} \\ &= 2.81 \times 10^{-10} \text{ m} \end{aligned}$$

The spacing between layers of ions in the crystal is $2.81 \times 10^{-10} \text{ m}$ or 2.81 \AA .

Sir Lawrence Bragg used the X-ray spectrometer designed by his father to analyse many crystals. Figure 13.33 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

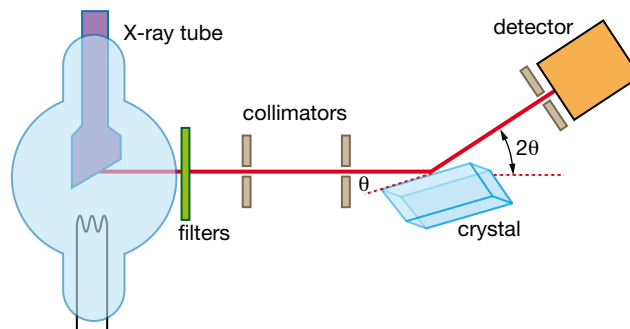


Figure 13.33 The Bragg spectrometer operates so that as the crystal rotates, the detector rotates through twice the angle. This set-up means that the detector remains in a position to register Bragg diffraction.

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.

Uses of X-ray diffraction

Synchrotron X-rays are the ideal tool for *X-ray diffraction*. The brightness and monochromaticity of the light allows high-contrast *X-ray diffraction* patterns to be created from crystallised samples. This process allows us to determine the structures of simple crystals and even to solve complex protein structures. By deriving the complex nature of biological macromolecules, scientists can better determine how they function. The use of three-dimensional structural information from X-ray diffraction is an essential key to better drug design.

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.

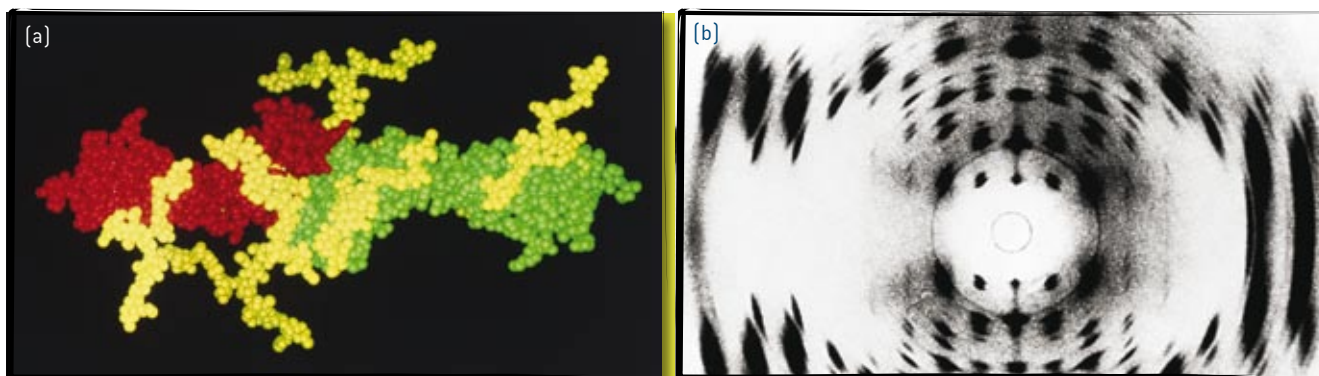


Figure 13.34 (a) X-ray diffraction can be used to analyse molecules as complex as the antigen glycoprotein. (b) The X-ray diffraction pattern produced on photographic film.

Great steps have been made in this field as synchrotron sources have become more advanced. X-ray sources are more intense, there is better beamline precision in delivering the radiation, electronic detectors used to record reflected intensities have improved and computers used for data analysis are now faster.

Electrons move in the space around the nucleus of the atom. It is the electrons of the sample atoms that diffract the X-rays. Because of this, we can calculate the average number of electrons per unit volume in a sample from its X-ray diffraction pattern.

Computers using purely digital data complete the analysis of the pattern by using calculations called Fourier transforms. Data is presented in the form of electron density maps. Each line on the diagram represents a contour of equal electron density. As a result, electrons are more closely crowded in areas where lines are bunched together. Study the electron density map of urea shown in Figure 13.35. Compare its shape with its structural formula. Looking at the electron density map for oxygen, can you suggest why it is made up of many more contours than the regions where you would find a hydrogen atom?

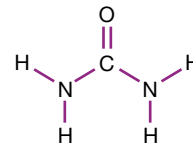
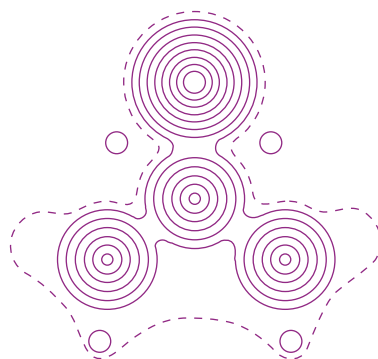


Figure 13.35 An electron density map and structural formula for urea (carbamide). The regions on the map corresponding to hydrogen atoms are particularly faint because there is only one outer-shell electron involved in diffraction processes.

Physics file

No doubt you are aware of the enormous amount of genetic information that can be extracted from the DNA analysis of a single strand of hair. Now there are more secrets your hair could reveal. An Australian researcher has observed that hair from women at a higher risk of developing breast cancer produces a different X-ray diffraction pattern compared with that of other women's hair. It is an exciting possibility for the future that a screening test could be as simple as examining a single hair with synchrotron light.



Figure 13.36 This is the three-dimensional structure of the RNA-producing enzyme polymerase. The structure of this enzyme was determined by diffraction methods from a synchrotron source. It is responsible for replicating the gene of the hepatitis C virus. Now that its structure is known, scientists are trying to develop a drug that will inhibit its function, and so stop the virus from replicating.

Most of our current understanding of the structure of complex bio-chemical molecules and materials is due to this process. The use of X-ray diffraction to discover the structure of crystals is called *X-ray crystallography*. This technique has been used to determine the structures of penicillin, vitamin B12 and haemoglobin. In 1953, Rosalind Franklin's high-quality X-ray diffraction images of DNA were essential to the landmark determination of the DNA molecule by Crick, Watson and Wilkins.

Physics file

The CSIRO (Commonwealth Scientific and Industrial Research Organisation) relies on powder diffraction techniques to analyse mineral specimens such as weathering products and clay minerals and industrial products like boiler scale, corrosion products, asbestos minerals and furnace slags and ashes.

Powder diffraction

The success of X-ray crystallography is due to the regularity of the structure of crystals. The process of X-ray diffraction can also be applied, with a polycrystalline powder as the target material. *Powder diffraction* is the method most commonly used to identify crystalline compounds, particularly those with a small grain size, typically less than 1 μm , such as soils, clay minerals and dusts.

A powder diffraction analysis can be conducted in a couple of ways. Most commonly, an analyser crystal selects the required wavelength of synchrotron light and directs a beam to the powdered sample. Because the sample is made up of a large number of individual crystals of different orientation, the spots produced in the diffraction pattern for an individual crystal then overlap and combine to form a number of concentric circles. These merge into a single, characteristic diffraction pattern for a substance.

An alternative method, which does not produce as high a resolution, may be used to record the entire pattern at the same time. In this case, the experiment is conducted at a constant angle, with a stationary energy-sensitive detector as shown in Figure 13.37. A thin layer of the sample powder is prepared on a piece of glass or plastic. This is inserted into the beam of X-ray photons. Diffracted beams can occur from all possible angles because of the random orientation of the sample. Once developed, the film reveals a pattern of vertical lines, spaced symmetrically either side of a bright central area. The spacings measuring from particular lines to the central region are a measure of the angle of diffraction.

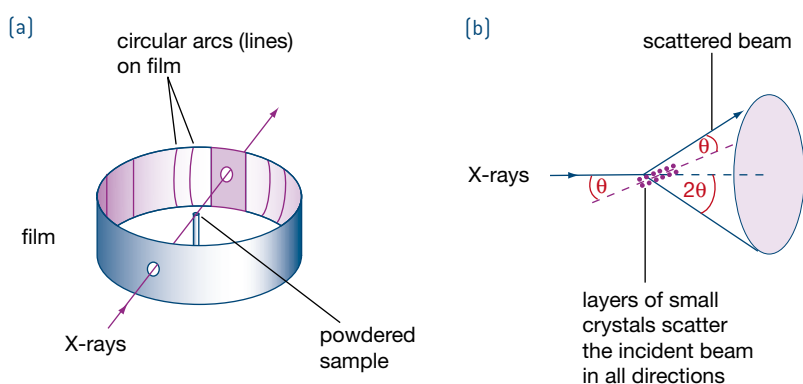


Figure 13.37 (a) A circular arc of X-ray film in an arrangement known as the Debye-Scherrer camera can be used to record the reflections on a strip of film. (b) Layers of small crystals scatter the incident photons through a wide range of angles.

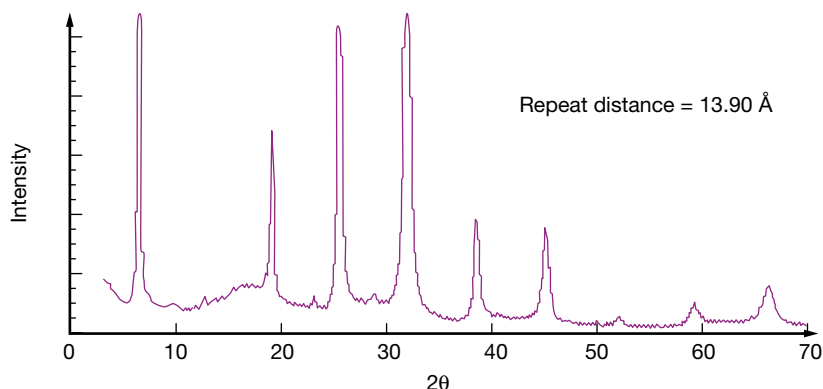


Figure 13.39 Either technique used for powder diffraction will record the intensity of diffraction peaks produced as a function of the diffracted angle θ or twice this value, 2θ . Computers are utilised to produce a graphical profile of the sample. To analyse data, these profiles can be compared with detailed profiles of previously examined substances. This image illustrates an example of the X-ray powder diffraction pattern obtained from a sample of vermiculite clay. The pattern of peaks formed in the profile may be used like a fingerprint to identify samples examined.

Physics file

Analysis of information linking diffraction intensity with diffraction angle makes it possible, using Bragg's law, to calculate values of d , or the parameters of the unit cell of which a compound is built. The unit cell is the basic structure of the compound, involving the particular atoms making up the structure, combined with the interatomic distances and angles between each. The International Centre for Diffraction Data (ICDD) currently has a database of over 60 000 known compounds. These can be cross-referenced with unknown samples to aid in their identification. This can be performed by using computer 'search and match' algorithms. If a complete match is not found, a further piece of data may be required to complete the identification.

Physics file

Forensic science has been identified as one of the priorities for usage of the Australian Synchrotron during its first few years of operation. The intensity and collimation of synchrotron light make the synchrotron an invaluable tool in successfully analysing micro-particles of evidence, such as skin, hair and paint, left at the scene of a crime. The FBI has employed such techniques using the beamlines of the Brookhaven Synchrotron in the USA.

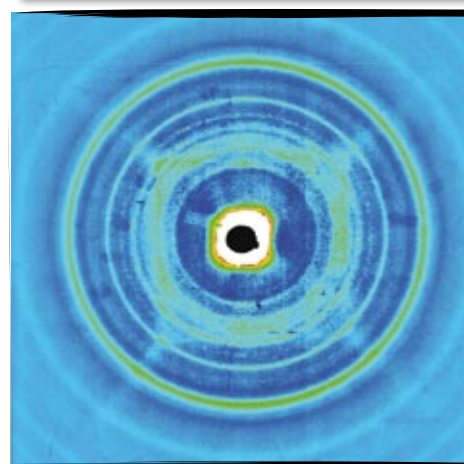


Figure 13.38 This is the powder diffraction pattern of a human heart valve that was produced at the Synchrotron Radiation Source at Daresbury. The concentric rings due to the averaging of diffraction patterns of the random orientation of many small crystals are clearly visible.

Relenza

The rise of synchrotron radiation sources has enabled scientists to access previously unknown details about the structure of molecules. In 1960, fewer than a dozen protein structures were identified down to the detail of individual atoms. At present, some 20 000 structures have been deposited in the Protein Data Bank, with more being revealed all the time. Armed with a better knowledge of molecular structures, scientists are poised to design drugs to inhibit the function of agents that cause disease.

The influenza virus relies on an enzyme called neuraminidase to spread it from cell to cell. If this could be inhibited, then the virus would not be able to spread through its host. In 1978, Dr Graeme Laver of the Australian National University's John Curtin School of Medical Research successfully crystallised neuraminidase. Dr Peter Colman and Dr Jose Varghese from CSIRO then spent years trying to unlock the structure of the protein. They used synchrotron sources from Tsukuba, Hamburg and Stanford to carry out X-ray diffraction experiments. They solved the structure of neuraminidase in 1983. Once the structure was revealed, they identified a particular section of the enzyme that appeared the same across all strains.

Professor Mark von Itzstein, at the Victorian College of Pharmacy at Monash University, worked with scientists from the firm Biota Holdings to design a drug that could lock into this particular section of neuraminidase and stop the virus from replicating. The drug they produced is called Relenza (zanamivir), the world's first anti-influenza drug. It has now been approved for therapeutic use in over 64 countries, and is marketed by GlaxoSmithKline.

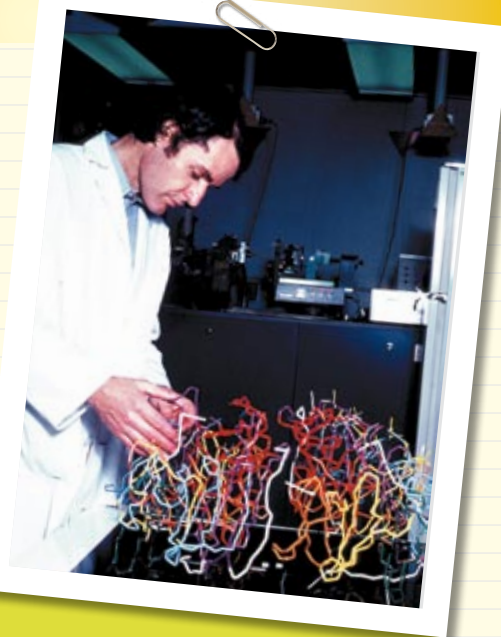


Figure 13.40 Professor Peter Colman working with a model of the influenza surface enzyme, neuraminidase.

Relenza is particularly used to treat members of the community in high-risk groups, such as the elderly and those with chronic respiratory problems, diabetes or cardiovascular disease. It is helping to prevent some of the approximately 1500 deaths that occur throughout Australia each year from influenza or related complications and has also been proven to be effective against the human strain of avian bird flu.



13.3 summary

Synchrotron radiation

- 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.
- This 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, is tunable and highly polarised.
- Synchrotrons are an ideal source of X-rays that can be used in a wide range of investigative applications. The X-rays are of very high intensity and are produced in a beam that has a lifetime of up to several hours.
- A collimated beam of X-rays incident upon a layer of atoms will be scattered. Bragg's law states that the beams will interfere constructively, producing maxima when the following relationship is satisfied: $2d\sin\theta = n\lambda$ where d is the distance between layers of atoms (m), θ is the angle the X-ray beam makes

with the surface, λ is the wavelength of incident X-ray photons (m) and n is the number of the maxima occurring, 1, 2, 3, etc.

- We can employ Bragg's law to analyse X-ray diffraction patterns. The structure of the unit cell defining a crystal can be identified through this process. This is called X-ray crystallography.

- Computers analysing X-ray diffraction data can be used to build an electron density map showing the locations of electrons in the macromolecule.
- Powder diffraction refers to the application of X-ray diffraction to a polycrystalline powder. This technique can be employed either by using a camera with circular film, or by scanning a detector to search through angles for diffracted beam maxima.



13.3 questions

Synchrotron radiation

- a** Explain the process by which synchrotron light is produced.

b List five characteristics of synchrotron light.
- Study Figure 13.24 to answer the following.

a What energy range (in eV) can synchrotron light have?

b To which range in wavelengths is this related?

c Explain why this range of wavelengths is useful for further synchrotron light applications.
- a** Consider Figure 13.25. Which range of energies of radiation is suitable for X-ray absorption and diffraction structure studies?

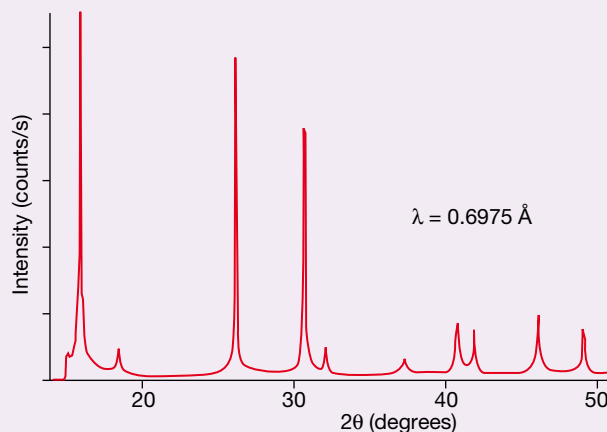
b The horizontal scale of the graph in Figure 13.25 reads from 100 keV to 1 eV. Reading from Figure 13.24, what types of electromagnetic radiation are found at 100 keV and 1 eV, respectively?
- High-energy electrons travelling in a curved path under the influence of a magnetic field will emit photons 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.
- List three differences between X-rays generated in an X-ray tube and X-rays produced in a synchrotron.
- In order to satisfy Bragg's law ($2d\sin\theta = n\lambda$), X-rays interacting with a sample crystal must:

A undergo destructive interference
B undergo constructive interference
C pass through the sample unchanged.
- A monochromatic X-ray strikes a crystal surface at a glancing angle of 31° . Detectors record a third-order maximum. Given that the spacing between layers of ions in the crystal is 5.6×10^{-9} m, calculate the wavelength of X-rays used.
- One benefit of using synchrotron light is that it is tunable, or in other words, it is possible to select electromagnetic radiation of a particular frequency for a specific function. Explain how a monochromator is able to achieve this selection of frequency by Bragg diffraction from a crystal.
- X-rays of wavelength 0.071 nm directed to a lithium chloride crystal produce a strong diffraction at an angle of 7.6° . Calculate the distance between adjacent planes of ions in the lithium chloride.
- This graph shows X-ray diffraction patterns created using a beam of collimated, monochromatic X-rays of wavelength 0.6975 \AA . Note that the angle between the detector and the outgoing beam is twice the glancing angle of incoming X-rays.

a Explain why you think a number of peaks have been recorded in this graph.

b Making an estimate of the angle shown to produce the first maxima (and remembering to halve the value from the graph to find θ), calculate an estimate of the spacing between layers of atoms in this crystal.



13.4 Scattering and beyond

You will recall that Max Planck (1858–1947) proposed in 1900 that light emitted by a hot object was delivered in small packets of energy, called quanta. Each *quantum* has energy $E = hf$, where h is Planck's constant ($6.63 \times 10^{-34} \text{ J s}$) and f is the frequency of light (Hz).

His ideas were not readily accepted because they implied that light behaved in some way like a particle. Planck also proposed that a single photon could only interact with a single electron at a time.

Using the wave equation:

$$c = f\lambda$$

where c is the speed of light (m s^{-1}) and λ is the wavelength (m)

the relationship $E = hf$ can be rewritten:

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

Once physicists discovered that light could carry energy in the same way as a particle carries energy, they realised that light may also possess momentum. This was later confirmed by Einstein when studying the photoelectric effect. The momentum, p , of a photon is said to be:

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

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 in Chapter 11. 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 13.41.

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.



The **PHOTOELECTRIC EQUATION** shows that ejected photoelectrons have a maximum kinetic energy: $E_k(\text{max}) = E_{\text{photon}} - W$, where $E_{\text{photon}} = hf$, which is the incident photon energy, and W is the work function, or energy binding the electron to the metal [$W = hf_0$].

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:

$$E_k(\text{max}) = eV_0$$

This value is therefore independent of the intensity of incident light and depends only on its frequency.

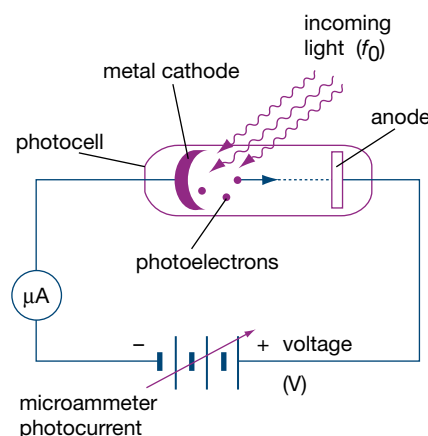


Figure 13.41 Incident light hitting the metal cathode will result in the emission of photoelectrons, provided that the frequency of the incoming light is sufficiently high. The voltage, V , can be varied continuously and also reversed by a switching arrangement.

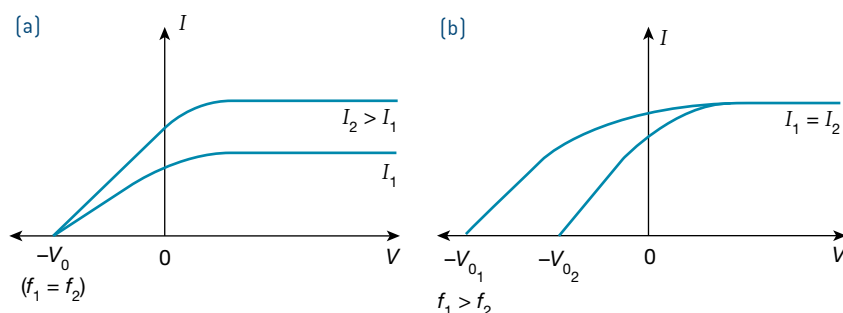


Figure 13.42 Photoelectric current plotted as a function of the applied voltage between the cathode and the anode. A standard monochromatic light source where $f > f_0$ shows that with a forward potential, every available photoelectron is included in the current. With a reverse potential, the number 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.

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

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.

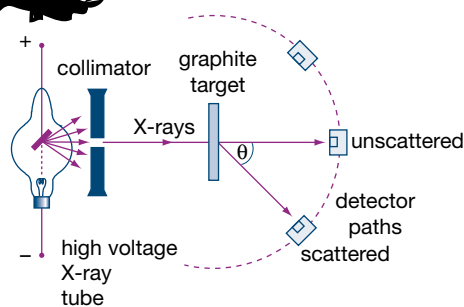


Figure 13.43 X-rays were made to collide with a graphite target. Some passed through unscattered and unaltered. Others were scattered and emerged with less energy than the unscattered X-rays.

Physics file

It can be shown, by considering the conservation of linear momentum, that the wavelength shift, $\Delta\lambda$, of the photons scattered in a Compton collision is:

$$\Delta\lambda = \lambda' - \lambda = \frac{h}{m_0 c} (1 - \cos\phi)$$

where λ' = the wavelength of scattered photons

λ = the wavelength of incident photons

m_0 = the rest mass of the electron

ϕ = the scattering angle

This wavelength shift depends upon the scattering angle, and not on the initial wavelength of incident photons. $\Delta\lambda$ varies from zero (if $\phi = 0$) when the photon is barely deflected to $\frac{2h}{m_0 c}$ for a head-on collision in which the photon bounces back from the electron. By considering the total energy of the system before and after the collision, we can say that:

initial photon energy = final photon energy + kinetic energy of electron emitted

or

$$\frac{hc}{\lambda} = \frac{hc}{\lambda + \Delta\lambda} + E_k$$

where E_k = the kinetic energy of the ejected electron

$\Delta\lambda$ = the change in the wavelength (or Compton shift) of the photon

Compton scattering of photons by electrons has played an influential role in promoting the wave-particle duality of electromagnetic radiation. It was important to the early development of quantum mechanics.

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 13.43.

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.

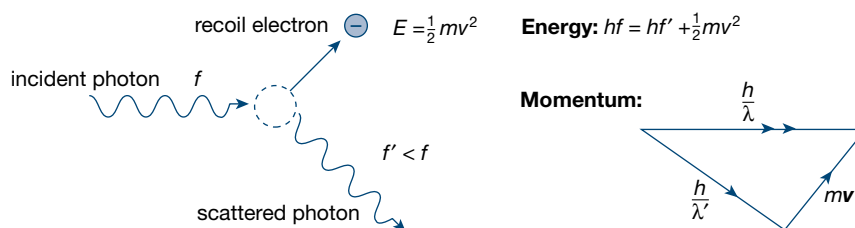


Figure 13.44 In the case of Compton scattering, energy and momentum are given up by some incident X-ray photons to electrons that are then ejected. These collisions are inelastic.

Compton equated the energy of a photon, $E = hf$, to Einstein's mass-energy equivalence relationship, $E = mc^2$, so that $E = mc^2 = hf$. So:

$$mc = \frac{hf}{c}$$

From here, he derived the equation for photon momentum (as previously stated) to be:

$$p = \frac{hf}{c} = \frac{h}{\lambda} \quad (\text{by substituting } c = f\lambda), \text{ where } f \text{ and } \lambda \text{ are the frequency and wavelength of the radiation.}$$



COMPTON SHIFT is the change in wavelength of photons resulting from inelastic collisions with electrons.

Experiments soon revealed that this Compton shift (or energy loss by the photons) was greater for larger scattering angles.

Because X-ray photons transfer some energy to electrons in a Compton collision, the collision is inelastic. As only a single electron is involved in each Compton collision at a specific point in space, there is just a single scattering wave. There are no favoured scattering angles as occur with Bragg diffraction, and no chance for interference. This produces the incoherent nature of Compton scattering, compared with the coherent and elastic Bragg diffraction.

Bragg diffraction, being coherent, yields specific spatial information about the structure of a sample. Compton scattering, being incoherent, does not give the same information. It is only sensitive to bulk properties, like momentum and density information. Its usefulness lies in the fact that it yields data complementary to that produced in elastic scattering methods.

Other scattering techniques

Diffuse scattering

If a sample with a regular array of molecules is studied by X-ray diffraction, then the refracted beams produce a characteristic pattern. If a sample is not well ordered, but of random orientation, then the *diffuse scattering* of X-rays produces a pattern of circular rings on the detector. Features of the sample are not so easily derived, unlike those of the regular crystalline samples.

The low intensity of scattered X-rays meant that in the 1950s and 1960s, scattering experiments would involve data collection over a number of hours or even days. Now these experiments are performed with brilliant synchrotron light sources, vastly shortening the time scales required and improving the resolution produced.

An added advantage has been the ability to use several powerful synchrotron techniques simultaneously. These include small and wide-angle scattering, absorption spectroscopy and X-ray diffraction methods. They can be used, for instance, to investigate the structure of new materials, aiding their development.

As previously discussed, X-ray diffraction from crystal structures will produce a two-dimensional pattern of clear Bragg peaks. These peaks are also broadened due to thermal diffuse scattering from the sample. Once this was seen as a nuisance, but physicists can now exploit this to gather information about lattice vibrations of the sample material.

To perform such experiments, a very precise knowledge of the wavelength and direction of the incoming X-ray beam is required. These characteristics may be compared with those of the reflected beam after a scattering experiment. This process can determine the nature of slightly disordered crystals which have a small variation to a regular lattice structure. Diffuse scattering between Bragg peaks can also be used to provide information about the movement of molecules in a crystal with respect to neighbouring molecules.

Small-angle scattering

Small-angle scattering is the name given to the techniques of small-angle neutron, X-ray and light scattering. Radiation is elastically scattered in each case and the resulting scattering pattern is analysed to give information about the sample.

These techniques are generally used to investigate non-crystalline biological macromolecular systems, for example, biological fibres, such as intact muscle. Scientists at the Daresbury Synchrotron Radiation Source have interpreted synchrotron diffraction data from the muscles of bony fish in terms of the so-called 'swinging crossbridge' model. This model describes how muscle filaments slide along each other to contract and relax the muscle. In this experiment, data were recorded over 5 ms time intervals while the tension in a relaxed muscle was gradually increased. This continued over a 200 ms timespan, producing some 40 sets of data about each stage of the contraction process. The advantage of using synchrotron radiation is that it allows direct observation of a change in a biological structure or other material, because data can be collected at high speed. This offers a new perspective, compared with the traditional approach of studying the sample after the change has occurred.

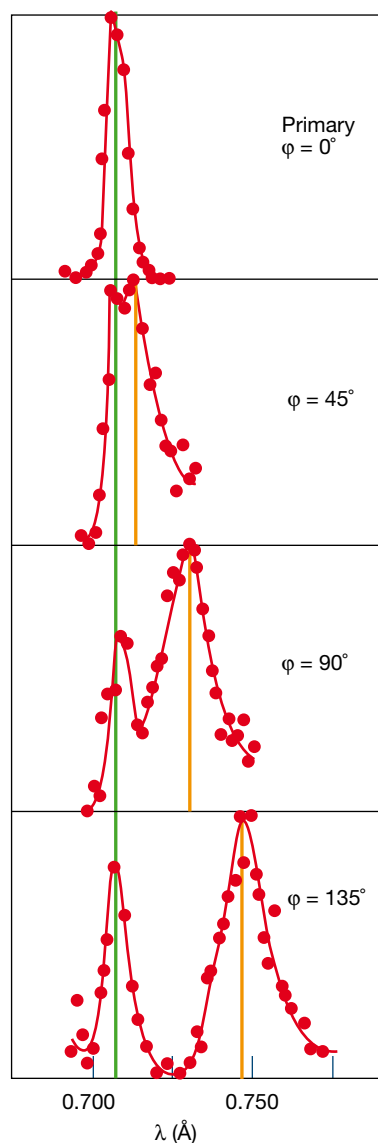


Figure 13.45 These peaks demonstrate Compton's experimental results. The green vertical line on the left represents λ and the orange line on the right is λ' .

Physics file

Small-angle scattering studies have been conducted by the Dow Chemical Company using synchrotron light. Scientists have completed real-time studies of engineering thermoplastics during tensile deformation. Other investigations include the study of the *in-situ* formation of polyurethane slabstock foams in real time during foam reactions.

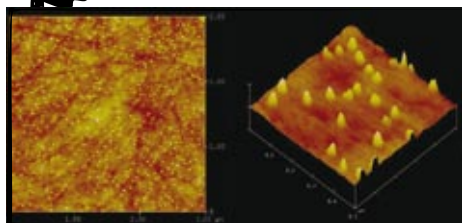


Figure 13.46 This image was produced using small angle X-ray scattering from the National Synchrotron Light Source at Brookhaven National Laboratory, New York. It shows the real-time formation of nano droplets of gallium on a sapphire surface and is utilised to better understand variations in the adsorption rate during this chemical process.

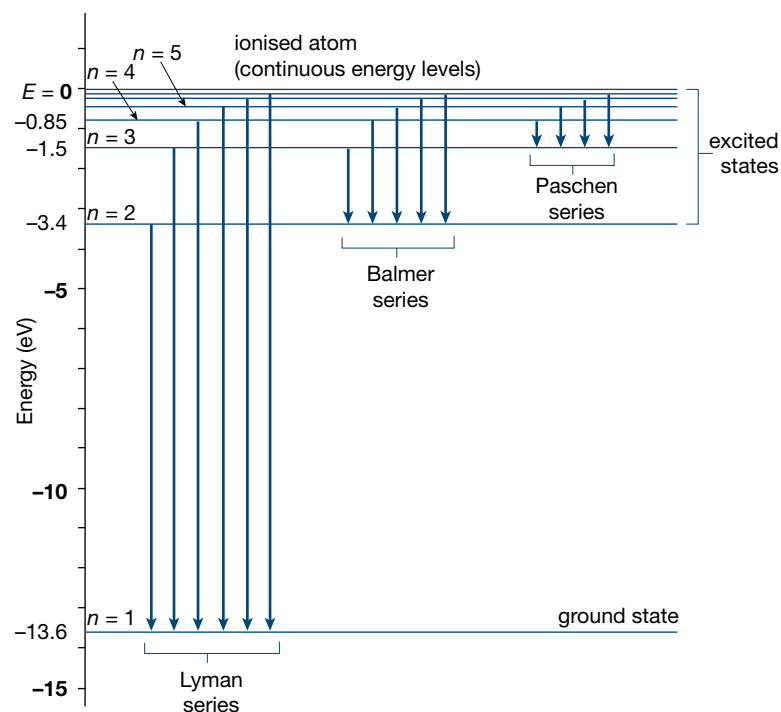
Figure 13.47 This energy level diagram for hydrogen shows that the ground state of the atom is 13.6 eV. Transitions from higher to lower energy levels will result in the emission of a photon of corresponding energy to the energy difference.

Physics file

When very high-energy photons from a synchrotron source are used, electron spin plays a role in the Compton collision. Scattering is dependent upon this spin orientation and becomes strongly magnetic. This scattering process has been used to study ferromagnetic systems, giving information about the spin momentum density of a sample. This is achieved by generating a magnetic Compton profile of a sample. A high-energy, circularly polarised, monochromatic X-ray beam is required for this process. The linearly polarised synchrotron beam can be converted to a circularly polarised beam, by using a phase plate or by a number of other methods. Recently, magnetic Compton profiles have been used to investigate a model made to simulate the emission of X-rays and gamma rays from the magnetosphere of the Sun.

Small-angle X-ray scattering can be performed on samples under conditions close to their normal physiological environment. Other studies have been made on crystal type, size and orientation in the mineralisation of bones and mollusc shells. Information gathered is of lower resolution than that for crystal structures, because the biological samples are usually only partially ordered, or of random orientation (amorphous). Using scattering techniques, we have gained a more detailed understanding of these biological structures and muscle function.

Spectroscopy



Electrons are found within specific energy levels, or shells, in an atom. An atom in its ground state is in the lowest energy state possible. An atom may accept bundles, or quanta, of energy that can raise it to higher energy levels, or to an excited state, by exciting electrons from inner to outer shells. Each energy level that atoms can occupy is given a principal quantum number, n , and may be denoted by a series of horizontal lines in a diagram, where $n = 1$ refers to the lowest energy level, or ground state. Figure 13.47 illustrates an energy level diagram for hydrogen. An electron moving from a higher state (E_2) to a lower state (E_1) releases light of frequency f , corresponding to this energy transition, so that $E_2 - E_1 = hf$.

Synchrotron light is suitable for a wide range of spectroscopic techniques. Line spectra are generally the result of transitions between atomic energy levels. Because high-energy X-ray photons possess much greater energy than photons of visible light, X-ray line spectra usually correspond to atomic transitions in the lowest energy levels. In such a case, X-ray photons displace an electron from an inner shell. An outer-shell electron will then fall into the hole created. As this occurs, an X-ray photon is released. This photon has energy equivalent to that of the transition between energy levels. Because the spacing between energy levels is specific for each element, the energy

of these fluorescent X-ray photons is characteristic of an element. Scientists use a technique called X-ray photoelectron spectroscopy (XPS) to study such photoelectron emissions.

In another useful spectroscopic technique, the degree to which a sample absorbs X-ray photons is measured as the energy of incoming radiation is increased. Generally, the proportion absorbed, called the absorption coefficient, declines with increasing X-ray energy. At the point at which X-ray energy levels correspond to the binding energy of an electron of the absorbing atom, we will naturally see a rapid rise in absorption energy. The energies at which this occurs are called absorption edges. At these values, electrons in the absorbing atom are ionised and ejected from the atom.

It could be assumed that once the X-ray energy increases beyond an absorption edge, the absorption coefficient will continue to gradually decrease. This does occur in the spectra of isolated atoms, such as the noble gases. However, the presence of other atoms in a molecule set up oscillations of the absorption coefficient near the absorption edge. This effect can be seen in Figure 13.48. The oscillations are characteristic of the surrounding molecule and as a result reveal structural data. Because this emission occurs as a result of the absorption of the beam, this method of detection acts as a measure of this absorption. This technique is called *extended X-ray absorption fine structure (EXAFS)*.

Substantial improvements have occurred in the development of high-intensity synchrotron light, beamline optics and detector sensitivity since the first synchrotron-based EXAFS experiments began in 1974. This technique is now widely used and can be applied to either crystalline or non-crystalline samples and even to liquids and gases. It is a tool that can give information about structural changes during a biological reaction. It has also been applied in the identification of the chemical state of contaminants and investigating the vulcanisation of rubber.

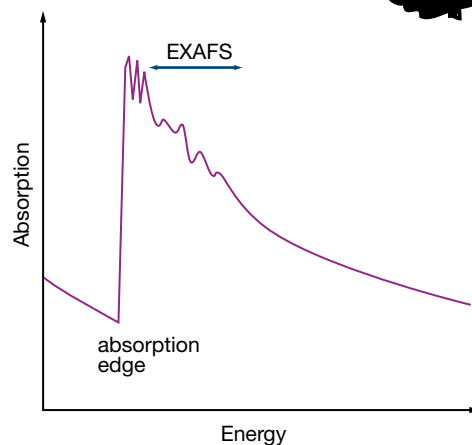


Figure 13.48 This curve shows the oscillations in absorption that occur just beyond an absorption edge. These oscillations are produced by interactions between ionised electrons and their structural spacing and configuration.



13.4 summary

Scattering and beyond

- Max Planck proposed in 1900 that light emitted by a hot object was not continuous, but was delivered in packets called quanta, each with energy $E = hf = \frac{hc}{\lambda}$.
- The momentum of a photon, p , is:

$$p = \frac{hf}{c} = \frac{h}{\lambda}$$
- X-rays can be elastically scattered in the case of Thomson scattering, or interact with electrons in the processes of the photoelectric effect, Compton (inelastic) scattering or pair production.
- 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 . Ejected photoelectrons have a maximum kinetic energy equal to:
 $E_k(\text{max}) = E_{\text{photon}} - W$, or $E_k(\text{max}) = hf - W$
 where W is the work function of the metal.
- X-ray photons can interact with electrons and transfer some energy, resulting in their emergence with a longer wavelength (and less energy), and the ejection of an electron. This is a Compton collision. These collisions are inelastic and the Compton shift in wavelength is dependent upon the scattering angle. Compton scattering and Compton magnetic scattering techniques provide information about the momentum density distribution of electrons in materials.
- Thomson scattering only occurs at lower energy levels of 100 keV or less.
- Synchrotron light is suitable for a wide range of spectroscopic techniques including the production of X-ray absorption spectra.



13.4 questions

Scattering and beyond

- 1 The three possible types of X-ray interactions with matter that involve energy transfer are:
 - A Bragg diffraction, Compton scattering, pair production
 - B Bragg diffraction, photoelectric effect, Compton scattering
 - C pair production, photoelectric effect, Compton scattering
 - D pair production, photoelectric effect, Bragg diffraction.
- 2 In the case of an interaction between an X-ray photon and an electron in the photoelectric effect, the ejected photoelectron accepts:
 - A none of the energy of the incident photon
 - B some of the energy of the incident photon
 - C all of the energy of the incident photon.
- 3 Light of which colour, red or blue, has the more energetic photons? Explain why.
- 4 Given that the threshold frequency for rubidium metal is 5.0×10^{14} Hz, calculate:
 - a the value of the work function of the metal
 - b the maximum velocity of ejected photoelectrons when the metal is illuminated by photons of frequency 2.8×10^{18} Hz.
- 5 Explain why Compton scattered X-ray photons emerge with their wavelength shifted compared with unscattered photons.
- 6 Diffuse scattering is used to examine partially ordered or amorphous structures. The analysis is only possible because of the brilliance of synchrotron beams. Explain the nature of information that can be gained by this technique.
- 7 This question is for extension purposes. The following is a list of analysis techniques that are possible using synchrotron radiation:
 - magnetic Compton scattering
 - powder diffraction
 - extended X-ray absorption fine structure
 - small-angle scattering
 - diffuse scattering.Select from this list the technique best suited to investigate each of the following:
 - a the electron configuration in inorganic crystals
 - b the structure of muscle fibres
 - c the nature of magnetic substances
 - d variation of lattice structure of a crystalline solid
 - e the atomic structure of a catalyst.
- 8 The threshold frequency for a particular metal is 2.8×10^{17} Hz. X-ray photons of frequency 5.8×10^{17} Hz are incident on the metal. Calculate:
 - a the wavelength of incident photons
 - b the incident photon energy
 - c the work function of the metal
 - d the maximum kinetic energy of the photoelectrons
 - e the maximum velocity of photoelectrons (assume the mass of an electron is 9.11×10^{-31} kg).
- 9 In his 1922 experiment in which he fired X-rays at a graphite sample, Arthur Compton found that the light that was scattered off the sample had a longer wavelength than that originally used.
 - a Did the X-rays fired towards the sample lose or gain energy in the collision?
 - b Suggest a second process that occurred during the collision that could account for the discrepancy in photon energy before and after the collision.
- 10
 - a Calculate the momentum of an X-ray photon of wavelength 520 \AA .
 - b Explain the importance of Compton's realisation that photons possess momentum in terms of Compton collisions.



chapter review

For each of the following questions, assume that:

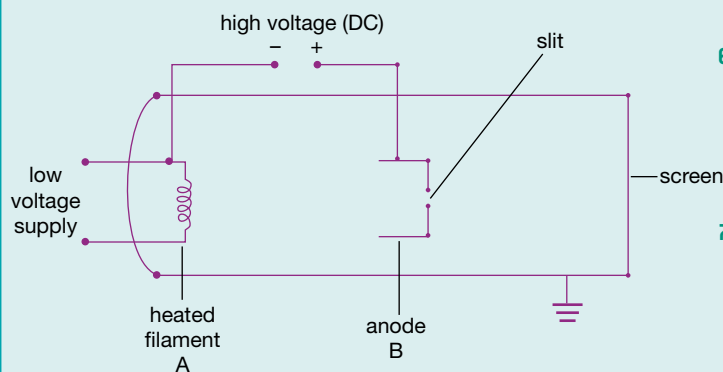
charge on an electron, $e = 1.6 \times 10^{-19} \text{ C}$

mass of an electron, $m = 9.11 \times 10^{-31} \text{ kg}$

Planck's constant, $h = 6.63 \times 10^{-34} \text{ J s}^{-1}$

Multiple-choice questions

- 1 Study the diagram of a simple cathode ray tube.

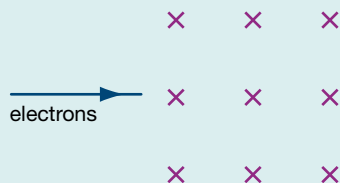


The source of electrons in this device is:

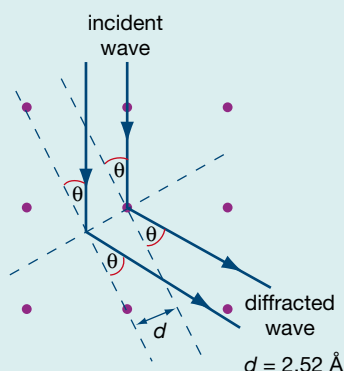
- A** the heated filament at A
B the positive anode at B
C the wires used in the circuit
D the screen used in the circuit.
- 2 A particular electron gun accelerates an electron across a potential difference of 15 kV, a distance of 12 cm between a pair of charged plates. The magnitude of the force acting on the electron is:
- A** $2 \times 10^{-17} \text{ N}$
B $2 \times 10^{-16} \text{ N}$
C $2 \times 10^{-14} \text{ N}$
D $2 \times 10^{-13} \text{ N}$
- 3 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.
- 4 In an electron gun, an electron is accelerated by a potential difference of 28 kV. The velocity with which the electron will exit the assembly is:
- A** $7.0 \times 10^7 \text{ m s}^{-1}$
B $4.9 \times 10^{15} \text{ m s}^{-1}$
C $9.9 \times 10^7 \text{ m s}^{-1}$
D $9.8 \times 10^{15} \text{ m s}^{-1}$
- 5 If the electron mentioned in Question 4 was accelerated a distance of 20 cm between a pair of charged parallel plates, then the size of the electric field acting is:
- A** 1.4 V m^{-1}
B $1.4 \times 10^3 \text{ V m}^{-1}$
C $1.4 \times 10^2 \text{ V m}^{-1}$
D $1.4 \times 10^5 \text{ V m}^{-1}$
- 6 Synchrotron light has 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.
- 7 If an electron travels through a magnetic field of strength 1.2 T with a velocity of $4.2 \times 10^6 \text{ m s}^{-1}$ then it will follow a path whose curvature of radius is:
- A** $1.6 \times 10^{-18} \text{ m}$
B $2.0 \times 10^{-5} \text{ m}$
C $6.1 \times 10^{17} \text{ m}$
D $5.0 \times 10^4 \text{ m}$
- 8 What is the energy in joules and electronvolts of a photon of wavelength $6.8 \times 10^{-11} \text{ m}$?
- A** $1.5 \times 10^{-52} \text{ J}, 9.4 \times 10^{-34} \text{ eV}$
B $2.9 \times 10^{-15} \text{ J}, 4.6 \times 10^{-34} \text{ eV}$
C $2.9 \times 10^{-15} \text{ J}, 1.8 \times 10^4 \text{ eV}$
D $1.5 \times 10^{-52} \text{ J}, 2.4 \times 10^{-71} \text{ eV}$
- 9 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.
- 10 X-ray photons of wavelength $9.8 \times 10^{-10} \text{ m}$ are incident on an aluminium sample material. Given that 4.2 eV is required to remove an electron, then the maximum kinetic energy of ejected electrons is:
- A** 1264 J
B 1273 J
C 1269 J
D 1260 J

Extended-answer questions

- 11** 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.

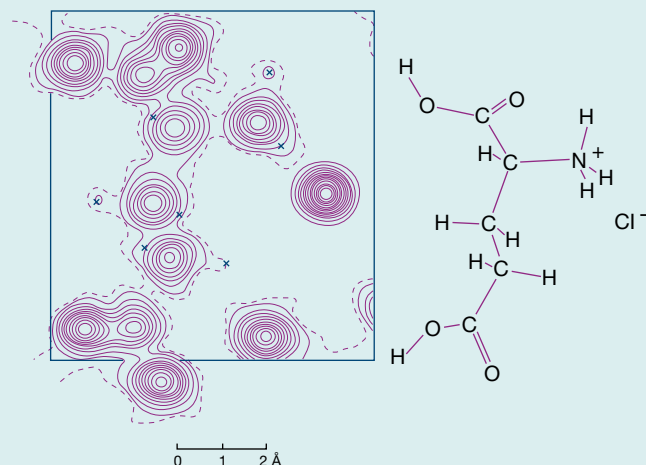


- 12** An electron beam travelling through a cathode ray tube is subjected to simultaneous electric and magnetic fields. The electrons emerge with no deflection. Given that the potential difference across the parallel plates X and Y is 3 kV, and that the applied magnetic field is of strength 1.6×10^{-3} T, calculate the distance between plates X and Y.
- 13** Synchrotron light is incident on a layer of sodium chloride atoms at a grazing angle as shown in the diagram. X-rays of 1.2 \AA are used in the experiment. If the first-order maximum is shown in the diagram, and the distance between layers of atoms is 2.52 \AA , then calculate the angles of subsequent diffracted beams.



The following information applies to questions 14 and 15.

The diagram shows an electron density map and structural formula of glutamic acid hydrochloride. The positions of the hydrogen atoms have not shown up clearly on the map and are estimated by crosses.



- 14** By referring to how the map was created, explain why the positions of the hydrogen atoms have not been revealed.
- 15** Find the region of greatest bunching of contours. Which atom do you think would be represented in this region? How did you come to this conclusion?
- 16** A band of X-rays of wavelengths $0.80\text{--}1.25 \text{ \AA}$ falls on a layer of ions with a glancing angle of 50° . Given that the spacing between layers is 2.62 \AA , would you expect to observe any more diffracted beams? Explain your answer.
- 17** 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 1.5×10^{-4} T. Given that they travel in an arc of radius 6 cm and that $\frac{q}{m} = 1.76 \times 10^{11} \text{ C kg}^{-1}$, calculate the speed of the electrons.
- 18** Two monochromatic X-ray beams are diffracted from a crystal. Ray P gives a third-order reflection maximum at a glancing angle of 60° from the face of the crystal. Ray Q, of wavelength 1.22 \AA , gives a first-order reflection maximum at an angle of 30° from the same crystal face. Find the wavelength of ray P.
- 19** An X-ray photon of wavelength $6.32 \times 10^{-12} \text{ m}$ collides with an electron in a Compton collision with a scattering angle of 90° . Would you expect the X-ray photon to emerge with a longer or shorter wavelength? Explain your answer.
- 20** Study the powder diffraction diagram of layered clay shown in Figure 13.39. Suggest how this could lead to information about its underlying chemical structure.



Photonics



Photonics has been defined as 'the science of using light to manipulate information and energy'. The study of photonics deals with all aspects of the generation, manipulation, transport, detection and use of light. It draws on theory covered in previous sections of the general study of physics, particularly an understanding of the nature of light. Applications of electronics and optics are also important.

Although photonics is a relatively new discipline in the field of physics, it has already played a significant role in shaping our modern life and will continue to influence our technological development over the coming decades.



by the end of this chapter

you will have covered material from the study of photonics, including:

- the nature and production of incoherent light
- sources of incoherent light, including LEDs
- the nature of coherent light and its production in lasers
- the nature, operation and limitations of optical fibres
- the use of optical fibres in telecommunications, imaging and sensor systems.



outcome

On completion of this chapter, you should be able to apply the photon and wave models of light to describe and explain the operation of different light sources and fibre optic wave-guides and analyse their domestic, scientific and industrial uses.

14

CHAPTER

14.1

Incoherent light sources

As you discovered in Year 11, optical radiation is just one form of electromagnetic radiation (EMR). It extends from infrared to ultraviolet. Visible radiation (i.e. EMR that our eye perceives as light) ranges from violet to red. In earlier chapters we found that light, and indeed any EMR, can be modelled either as a continuous wave or as a stream of photons, each with a discrete amount of energy:

$$E_{\text{photon}} = hf = \frac{hc}{\lambda}$$

where f = the frequency of the light photon

λ = the wavelength of the light

c = the speed of light in a vacuum ($2.998 \times 10^8 \text{ m s}^{-1}$)

h = Planck's constant ($6.626 \times 10^{-34} \text{ J s}$).

Coherence and incoherence

EMR can be explained as being produced whenever electric charge accelerates. For example, in a radio transmitter, an oscillator generates the amplified radio signal that drives the electrons so that they vibrate back and forth along the transmitter aerial. All the electrons oscillate with exactly the same frequency and phase as the energising signal. Their motion is, in effect, synchronised by the energising signal. As the oscillating electrons are undergoing continual acceleration, their movement generates the radio waves that radiate out from the aerial. The individual electromagnetic (EM) waves generated from all the electron oscillations add up to form one synchronised EM wave. Because of this synchronisation, the radio wave is said to be *coherent*. The electrons in the aerial of the radio receiver in your mobile phone are set into oscillation—all with the same frequency and phase—by these radio waves, and hence generate an electrical signal of the same frequency that is converted back to a copy of the original audio signal.



Points in the path of a wave motion, such as EMR, are said to be in phase if the displacement and velocity at those points are exactly the same at any instant.

Because the radio wave is coherent, interference effects can often be observed. For example, the radio signal from a transmitter can sometimes reflect off a building and interfere with the signal arriving at the aerial directly. In this case, destructive or constructive interference can occur and the radio signal may fade or grow as the radio receiver is moved from one location to another.

Now consider the hypothetical situation where instead of one big oscillator with a single frequency driving one big aerial, we use a thousand independent small oscillators, each with a completely different frequency and phase with respect to each other. If each of these small oscillators were to energise its own small aerial, then the radio waves that each would produce would also have different frequencies and phases (i.e. the waves would not be synchronised). In this case, the sum of all these waves would be a complex waveform with a large spread of frequency components and a constantly changing resultant phase. This lack of synchronisation would result in a wave that is said to be *incoherent*. Incoherent radio waves are usually not

See Physics in action 'Light—an electromagnetic wave', page 390.

Physics file

While waves must have the same frequency and be in phase to be coherent, they do *not* have to have the same amplitude. Imagine it as similar to people being in step. You don't have to be stepping the same distance—just in time.



Figure 14.1 A transmitter and a receiver of radio waves.

very useful, as the constantly changing phase would, on average, result only in a constant 'noise' signal at the receiver.

Earlier in this course we looked at Young's famous experiment in which he showed that the diffracted light from two very close slits creates the characteristic 'two source' interference pattern of light and dark bands. What would happen if we were to replace the two slits with two fine wire filaments an equal distance apart? It will probably not be a surprise to you that we would see no interference pattern. The key to Young's experiment was that the two slits were illuminated by the same source and therefore the light from the two slits, coming from the same source, was always in phase. On the other hand, there would be no constant phase relationship between the light from two wire filaments, and so any pattern would be very fleeting. As a result we would just see a uniform spread of light. As with the radio waves discussed above, the light waves from the two filaments would be incoherent.

Continuous wide-spectrum light sources

All atoms in matter are in continuous motion due to their thermal energy. Because of the complex nature of the interaction between the many atoms in a crystal, there are very many different modes of oscillation, and the natural frequencies of these oscillations have a very broad distribution.

Considered from a wave perspective, the oscillating atoms in a crystal consist of accelerating charges that act like tiny incoherent antennas, giving rise to the emission of electromagnetic radiation (called *thermal radiation*). Thermal radiation is always emitted from any matter with a temperature above absolute zero (-273.15°C).

A thermal radiator at low temperature has charges that, on average, oscillate at lower frequencies and hence it emits radiation that, on average, is dominated by longer wavelengths. A thermal radiator at high temperature has charges that oscillate at higher frequencies and hence emits radiation that, on average, is dominated by shorter wavelengths. Because there is a large range of possible frequencies for the oscillating charges in a body at a particular temperature, there is a continuous distribution of wavelengths over a wide range in the electromagnetic radiation being emitted. The electromagnetic radiation intensity distribution for any radiator is called its *electromagnetic spectrum*. In a sense, this incoherent thermal radiation is similar to the incoherent radio waves discussed earlier in this section.

The Sun, incandescent light bulbs and candle flames are all examples of incoherent thermal radiators producing electromagnetic radiation including visible-light radiation. Actually, any body with a temperature above absolute zero is a thermal radiator.

The electromagnetic spectrum generated by any ideal thermal radiator always has the same basic shape when we plot the intensity of the radiation versus its wavelength. The electromagnetic spectra from several thermal radiators of different temperatures are shown in Figure 14.2. As you can see from the diagram, the intensity distribution of thermal radiation usually extends from the ultraviolet, through the visible, to the infrared part of the EM spectrum. The distribution for each thermal radiator is somewhat asymmetric and has one single peak. The wavelength at which the peak of the distribution occurs varies, according to the surface temperature of

Physics file

While incoherent radio waves are generally not very useful, the 'spread spectrum' transmission system is one form of incoherent radio-wave transmission that is useful. Spread spectrum transmission is used in the global positioning system (GPS), developed by the US Department of Defence. GPS uses a system of 21 satellites that transmit a coded pseudo-random complex waveform (with many frequency components). A GPS receiver on Earth can decode the signal from four or more satellites to determine its precise three-dimensional position to an accuracy of a few metres. The reason why the military uses a spread spectrum system for their GPS is that it is very difficult to jam, since the coded signal is transmitted over a wide spread of frequencies. Conventional (coherent) radio waves that use only one single carrier frequency, on the other hand, are very easy to jam.



PRACTICAL ACTIVITY 42

Young's interference experiment

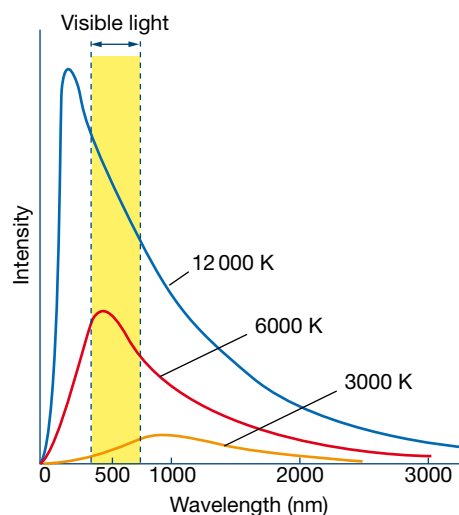


Figure 14.2 Electromagnetic spectra from several thermal radiators with different surface temperatures.

More on energy efficiency and conservation can be found in *Heinemann Physics 11* Chapter 14, 'Investigations: Sustainable energy sources'.

the radiation-emitting body. High-temperature thermal radiators have the peak of their EM emission at a higher frequency and shorter wavelength (i.e. towards the UV end of the spectrum) than low temperature radiators.

Incandescent light bulbs

The tungsten filament of an incandescent light bulb sits in an atmosphere of inert gases (argon and nitrogen) encased in a glass globe. The inert gases are used instead of air so that the reactive tungsten will not burn when heated to temperatures of about 2500 K. The inert gases also reduce the evaporation rate of the filament and therefore extend the incandescent light bulb's life. Eventually, evaporation causes a localised thinning at one point in the filament. This means the localised resistance and hence power dissipation increases at this point. This causes even more evaporation, and so on, until the filament breaks at this weak point. If you look at an incandescent light bulb that has 'blown' in this way, you can often see a faint deposit of evaporated tungsten powder on the inside of the globe.

About 80% of the electric power dissipated in a 60 W incandescent light bulb is converted to thermal radiation; the remainder is lost through conduction and convection. Only about 12% of the thermal EM radiation is produced in the visible part of the spectrum. So an incandescent light bulb has a very low efficiency (about 12% of 80% = 5%) for converting electrical energy into light energy. Other types of light sources, including fluorescent lamps and LEDs, are much more energy efficient.

Tungsten-halogen bulbs

Tungsten-halogen incandescent bulbs, also known as quartz-halogen bulbs, are more efficient than standard incandescent bulbs because they are designed to operate at a much higher filament temperature (3300 K), very close to the melting point of tungsten (3695 K). Tungsten-halogen bulbs can be found in many down-lights in houses and shops and are also increasingly used in the headlights of expensive cars.

The higher temperature of the tungsten filament means that the peak in the thermal radiator's electromagnetic spectrum occurs at a shorter wavelength, and hence more of the thermal emission appears in the visible part of the spectrum. To cope with the elevated operating temperature, the globe encasing the tungsten filament is usually made of quartz rather than glass.

So why doesn't the tungsten filament of a tungsten-halogen bulb evaporate away at a fast rate? The inert atmosphere surrounding the filament is similar to that in the standard bulb, but it also contains an active element (a halogen, which is usually either bromine or iodine). Any evaporated tungsten atoms form a semistable tungsten-halogen compound when they are close to the surface of the relatively cool quartz bulb. These compounds float around inside the bulb until they come close to the filament again. The extreme heat of the filament then breaks the compounds apart and the tungsten atoms are redeposited onto the filament. This means that tungsten-halogen bulbs last much longer than conventional incandescent bulbs. They also emit 10–15% more light for the same input electrical energy.

Candles, stars and Wien's law

Although the mathematical description of the shape of the spectral distribution curve for an ideal thermal radiator is quite complex, there is a very simple inverse relationship between the peak wavelength and absolute temperature. This relationship is known as *Wien's law*.

Wien's law allows us to determine the approximate temperature of a radiating body from its colour. For example, astronomers use special colour filters and very sensitive photon detectors to accurately calculate the peak wavelength of the light emitted from stars. The peak wavelength can then be used to determine the surface temperature of the star. Astronomers often refer to hot 'bluish' stars or cool 'reddish' stars.

The Sun is also an example of a thermal radiator. Its surface temperature is approximately 5500–6000 K. Using Wien's law, this temperature can be calculated from the wavelengths of visible light emitted by the Sun, which are around 540–480 nm (i.e. in the yellow-green part of the spectrum). The human eye has evolved to be most sensitive to wavelengths in roughly the same part of the spectrum. While this seems to make good evolutionary sense, other species—perhaps surprisingly—have evolved with greater sensitivities in other near-visible parts of the spectrum.

Much cooler bodies also emit thermal radiation. A body whose surface temperature is around room temperature (approximately 300 K) emits EM radiation peaked at approximately 10 μm (in the far infrared part of the spectrum), well beyond the sensitivity of the human eye. Humans have a body temperature slightly above room temperature and hence emit EM radiation with a peak a little below 10 μm . *Thermal imaging* cameras can detect small variations in infrared intensity. They can be used to image objects whose temperature is slightly above the ambient level, even in total darkness (i.e. when there is no visible light). The infrared image is electrically converted into a visible image or *thermograph*.



Figure 14.4 Most objects at everyday temperatures emit some radiation in the infrared wavelengths. A thermograph showing the infrared radiation can be used to find a lost person, indicate where energy is being lost from a building, predict weather patterns or assist in medical diagnosis.

Physics file

Wien's law can be written as:

$$\lambda_{\text{max}} = \frac{\alpha}{T}$$

where λ_{max} = the peak wavelength (m)

T = the absolute temperature (K)

α = the Wien constant, which (in SI units) has a value of $2.898 \times 10^{-3} \text{ m K}$.



Figure 14.3 The variation in colour of a candle flame is really a temperature profile, since the wavelength of the emitted light is related to the temperature of the burning gas.

Candles in space

Candle flames behave differently if they are in a weightless environment or in the apparent weightlessness of an orbiting space shuttle. All the air molecules in the space shuttle cabin are in free-fall, and hence the air lacks the buoyant convection that normally plays an important role in maintaining and shaping a flame on Earth.

Because there is no buoyant convection, the transport of combustion products and oxygen occurs by the much slower process of molecular diffusion. This diffusion occurs when there is a high concentration of combustion products and a low concentration of oxygen close to the flame and a high concentration of oxygen farther away from the flame. The combustion products migrate away from the flame and the oxygen migrates towards the flame. The diffusive transport rates are much lower than the transport rates due to natural convection on Earth.

As a result, a flame in an orbiting shuttle will often appear to burn less vigorously but with a bluer colour (hotter) than a flame on Earth. It will also assume a spherical shape that diffuses equally in all directions (see Figure 14.5), rather than the more elongated shape that is characteristic of flames on Earth.



Figure 14.5 A candle flame in the apparent weightlessness of an orbiting space shuttle.

Narrow-spectrum discrete-light sources

In Unit 4, Area of study 2, we found that when an atomic electron drops from a higher quantised energy level to another, lower, energy level, a photon is released. The energy of the emitted photon is also quantised and is equal to the difference between the discrete energies of the two states:

$$E_{\text{photon}} = \Delta E = E_2 - E_1$$

where E_{photon} = the energy of the released photon

E_2 = the higher quantised energy level of the atom

E_1 = the lower energy level of the atom.

When atomic electrons in essentially independent atoms, like those found in a gas, are excited to higher atomic energy levels (e.g. by collisions with energetic free electrons), usually there are only a few well-spaced levels that are excited. With thermal excitation in a lattice, on the other hand, there are many different modes of oscillation. These result in a continuous wide spectrum of incoherent thermal radiation. The excitation of isolated atoms, however, results in incoherent optical radiation emitted at just a few discrete frequencies or wavelengths. These are the spectral lines we see through a spectroscope when various elements are burnt in a Bunsen flame.

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.

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 bluish-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.

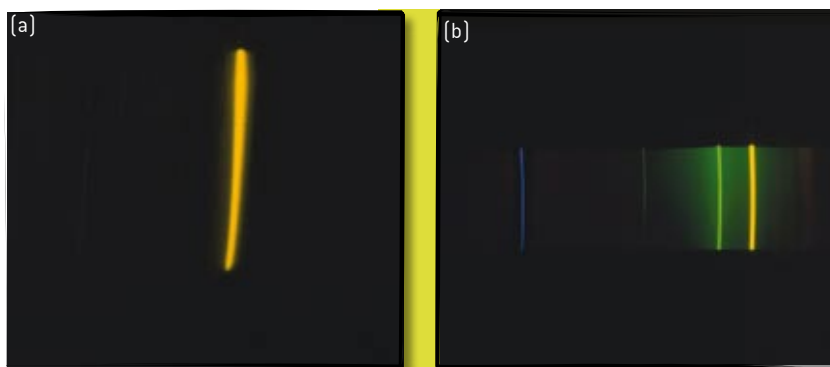


Figure 14.6 Emission spectra of (a) sodium and (b) mercury. The emission spectrum of a sodium- or mercury-vapour lamp consists of a large number of discrete frequencies, some of which lie in the visible region of the spectrum. Note the strong yellow line which gives sodium light its characteristic colour.

Fluorescent lamps

The fluorescent lamp is another variation on the mercury-vapour lamp. In a fluorescent lamp, the mixture of mercury vapour and argon gas is at a much lower pressure than in a standard mercury-vapour lamp. As a result, the distance between collisions for the accelerating electrons is much larger and the kinetic energy gained by the electrons is much greater. This means that electrons can be excited to much higher energy states in the mercury atoms. The de-excitation of these higher energy states generates photons of higher energy and hence shorter wavelengths (in the ultraviolet part of the optical spectrum). To produce visible radiation, the inside of the fluorescent tube is coated with a phosphor. The phosphor absorbs the higher-energy shorter-wavelength UV photons and then re-emits lower-energy longer-wavelength photons in the visible part of the spectrum.

Physics file

Increasingly, modern 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.

Physics file

Different fluorescent lamps have different phosphor coatings that can generate light of different colours. For example, normal fluorescent lamps for the home have a phosphor that emits whitish-blue light, whereas the fluorescent lamps in supermarket meat displays use a phosphor that emits pinkish-white light. The special 'black light' fluorescent lamps in banks have no phosphor coating and emit low-intensity UV radiation. These lamps are used to illuminate 'invisible' signatures in bankbooks that have been written with phosphorescent ink. Their radiation also makes some dyes in clothes light up, an effect you may have enjoyed when visiting a nightclub.

Electroluminescence

The forward-biased pn junction of a semiconductor can also generate incoherent narrow-spectrum light. The forward bias decreases the effective potential barrier across the pn junction and allows a large current to flow across the junction. This means that a large number of electrons are swept from the n-type region through the depletion region into the p-type region, where they can fill one of the many holes. This is called recombination. Similarly, holes migrate from the p-type region into the n-type region, where they can recombine with electrons. Hence, the current that is injected across the forward-biased junction gives rise to recombination and a subsequent release of energy. In certain semiconductors (such as gallium arsenide), this energy is released in the form of photons. This process is called *electroluminescence* and it is the mechanism that allows the light-emitting diode (LED) to generate optical radiation.

In terms of energy levels, semiconductors have two discrete bands of extremely closely spaced levels—so close that each energy level in a particular band is indistinguishable from any other. The lower band, which is generally full of electrons, is called the *valence band* and the upper band, which is generally empty of electrons, is called the *conduction band*. When a loosely bound electron gains enough energy, it can break free of the lattice. In energy level terms: the electron is raised into the conduction band. Any electron in the conduction band is essentially free to move about in the lattice. If the pn junction of a semiconductor is forward biased, these free electrons are attracted across the depletion layer, where they can easily recombine with holes. This corresponds to an electron falling from the conduction band to fill a hole (the absence of an electron) in the valence band. Since the electron is transferring to a lower energy state in the lattice system, there is a release of energy—a photon of optical radiation, in the case of a light-emitting diode. The energy of the photon is equal to the difference in energy between the conduction (E_c) and valence bands (E_v), termed the *band gap energy* (E_g) for that particular semiconductor. That is:

$$E_{\text{photon}} = E_g = E_c - E_v = hf = \frac{hc}{\lambda}$$

where λ is the wavelength of the emitted photon.

Different semiconductor materials have different bonds holding the atoms of the lattice together, and hence the energy associated with recombination (i.e. the re-forming of incomplete bonds) varies between different semiconductor materials. In energy level terms, we say that different semiconductor materials have different band gaps. This means that the photons emitted from different semiconductor materials have different peak emission wavelengths.

The mechanism of the LED has been described in detail in Chapter 5 'Introducing photonics'.

For more on semiconductors, see Physics in action 'Inside a diode' on page 131.

Energy bands in solids—chemistry is just physics after all!

A single independent atom, as found in a gas such as hydrogen, has discrete, well-spaced atomic energy levels (corresponding to electrons in different discrete orbitals). For an atom in its ground state, all the lowest energy levels are filled and the higher energy levels are vacant.

A system of several identical atoms bound together (as a molecule) still has the same discrete, well-spaced energy levels, but now each level divides into a number of closely spaced states. This occurs because of the nature of the complex forces acting between all the atoms in the molecule. The more atoms there are in the system, the more closely spaced states exist for each level in the system. If a large number of atoms come together to form a solid, each of the original atomic energy levels becomes a band. The energy levels in each band are so close together that they can seem continuous. That is why the thermal radiation spectrum of solids, discussed earlier, appears continuous.

In the case of a *good conductor*, the highest energy band is only partially filled with electrons, or else two bands overlap so that unoccupied states are available. When a potential difference is applied across the conductor, electrons can respond by increasing their energy, since there are plenty of unoccupied states of slightly higher energy available. Sodium, for example, is a good conductor because the highest energy 3s band (check with your chemistry text!) is only half full.

On the other hand, a *good insulator* is a material where the highest energy band, referred to as the valence band, is completely filled. The next highest band, the conduction band, is separated by an energy or *band gap*. For insulators, this energy gap (E_g) is around 5–10 eV. At room temperatures almost no electrons receive sufficient energy to reach the conduction band. When a potential difference is applied to the

material, no unoccupied states are available; therefore, no current flows and the material is a good insulator.

In a *semiconductor* lattice, the valence and conduction bands are separated by a much smaller energy gap, typically in the order of 1 eV. At room temperatures a few electrons can acquire enough energy to ‘jump’ the band gap, and a very small current may flow when a potential difference is applied. At higher temperatures, more electrons have sufficient energy and a higher current flows.

But that is not all that’s going on in a semiconductor. When a potential difference is applied, the few electrons in the conduction band move towards the positive electrode. Electrons in the valence band try to do the same thing, and a few can, filling the ‘holes’ left empty by electrons reaching the conduction band. Each electron in the valence band filling a hole in this way leaves behind a hole, so that holes tend to drift towards the negative electrode. Electrons tend to accumulate on one side and holes on the other, and hence there is a flow of current.

In a single atom, electrons can only exist in one of the discrete energy levels. For example, an electron can exist in the last filled energy level or in the first unfilled energy level; it cannot exist halfway in between. Similarly, in a lattice, electrons can only exist somewhere in the filled valence band or somewhere in the filled conduction band, but not anywhere in between. The region in between the top of the valence band and the bottom of the conduction band is called the ‘forbidden band gap’. In an LED semiconductor lattice, the energy difference between the top of the valence band and the bottom of the conduction band is called the band gap energy (E_g) and corresponds to the energy of the emitted photons.

Worked example 14.1A

At what wavelength will an LED radiate if it is made from a material with a gap band energy, E_g , of 1.84 eV?

Solution

$$\text{Since } E_g = hf = \frac{hc}{\lambda}$$

$$\text{Then } \lambda = \frac{hc}{E_g}$$

$$\begin{aligned} \text{So } \lambda &= \frac{6.63 \times 10^{-34} \times 3.00 \times 10^{-8}}{320 \times 10^{-9} \times 1.6 \times 10^{-19}} \\ &= 6.74 \times 10^{-7} \text{ or } 674 \text{ nm [a red LED]} \end{aligned}$$

Worked example 14.1B

The conductivity of a certain semiconductor has been found to increase when light of wavelength 320 nm or shorter is incident upon it.

- Why would this occur?
- What is the E_g for this semiconductor in eV?



Figure 14.7 After starting out as simple red 'power on' indicators, LEDs are now produced in virtually every visible and near-visible colour. Everything from high-intensity torches to lighting for fibre-optic systems is catered for.

Solution

- a** For the conductivity to increase, electrons must be passing from the valence band to the conduction band.
- b** Since $E_g = hf = \frac{hc}{\lambda}$
- $$= \frac{6.63 \times 10^{-34} \times 3.00 \times 10^8}{320 \times 10^{-9} \times 1.6 \times 10^{-19}}$$
- $$= 3.88 \text{ eV}$$

LEDs

Table 14.1 lists the peak wavelengths that LEDs have been designed to emit and the various semiconductor materials used. By varying the percentages of the different elements making up the LED, the wavelength can be varied to suit particular applications. While just a few years ago LEDs came only in infrared, red, yellow and a sort of yellowish green, they now come in just about every colour for which a commercial application could be imagined.

Table 14.1 Wavelengths and compositions of LEDs

Colour	Wavelengths (nm)	Semiconductor materials
White		InGaN + phosphor
Ultraviolet	370–390	GaN
Blue	430	GaN + SiC
Blue	450 and 473	InGaN
Turquoise	495–505	InGaN
Green	525	InGaN
Yellow-green	555–575	GaAsP and related
Yellow	585–595	
Amber	595–605	
Orange	605–620	
Orange-red	620–635	
Red	640–700	
Infrared	700–1300	

White-light-emitting diodes (WLEDs) are now being used in many lighting applications. True white light contains all colours and cannot be directly created by a single LED. WLEDs are not truly white at all—just very good approximations. One common form of 'white' LED uses a gallium nitride blue LED coated with a yellow phosphor material that, when excited by the blue LED light, emits a broad-range spectrum which added to the blue emission makes a fairly white light. The light has a slightly dominant blue component and is similar in colour to a mercury-vapour street lamp. Other WLEDs are made up of three LEDs placed very close to each other on the one head. The three LEDs emit red, blue and green light respectively and their intensity is adjusted in the manufacturing process to produce a close approximation to white light.

By varying the percentage of elements making up an InGaAsP LED, the wavelength can be varied from 1000 to 1550 nm: the infrared wavelengths used in modern fibre-optic communication systems.



14.1 summary

Incoherent light sources

- In coherent electromagnetic (EM) radiation, the component radiations keep the same phase relationship to each other. Interference effects can be observed.
- In incoherent EM radiation, the component radiations vary in phase. The resulting wave is complex.
- In a wide-spectrum EM radiation source, the complex nature of possible interactions produces a broad spectrum. Examples include the Sun, candle flames and incandescent light bulbs.
- The peak wavelength of a wide-spectrum thermal radiator is given by Wien's law:

$$\lambda_{\max} = \frac{\alpha}{T}$$

where λ_{\max} = the peak wavelength (m), T is the temperature of the radiator (K) and α is the Wien constant, 2.898×10^{-3} m K.

- LEDs produce incoherent light in a process called electroluminescence, as electrons drop from the conduction band to the valence band. The wavelength of the radiation produced depends on the energy gap between bands:

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

where E_g is the energy of the photon, h is Planck's constant (6.63×10^{-34} J s), f is the frequency of the radiation, c is the speed of light (3×10^8 m s⁻¹) and λ is the wavelength of the radiation.



14.1 questions

Incoherent light sources

- 1 Light from a single source falls on two slits 0.5 mm apart. An interference pattern appears on a screen 2.5 m away. The slits are now replaced by two tiny light globes. No interference pattern is observed. Explain this.
- 2 In relation to radiation such as EMR that exhibits wave-like behaviour, what is meant by the term 'in phase'?
 - A The amplitude of the waves is the same at all points along the wave.
 - B The displacement of any point on the wave is the same.
 - C The velocity of any point on the wave is the same.
 - D Both the velocity and the displacement of any point on the wave are the same.
- 3 The relative surface temperatures of stars many light-years away can be determined on the basis of the colour of the light they emit. Explain how this is possible.
- 4 A particular material has its valence band completely filled. The next highest band is separated from the valence band by an energy gap of 6.3 eV. This material will be:
 - A a conductor
 - B an insulator
 - C a semiconductor.
- 5 A semiconductor is exposed to light of slowly shortening wavelength. It begins to conduct when the wavelength has fallen to 680 nm. Calculate the size of the energy gap.
- 6 A TV remote control operates using an LED that emits infrared light. The detector on the TV is covered by a filter to stop it from detecting visible light. The filter only lets wavelengths through that it is responsive to. If silicon has an energy gap of 1.14 eV, could it make an effective filter? (It should not be responsive to visible wavelengths.)
- 7 The Sun is a source of incoherent light. Most of the Sun's radiation is of wavelengths shorter than 1000 nm. For the material of a solar cell to be able to absorb this light, what energy gap should the material have?
- 8 What colour will an LED radiate if it is made from a material with an energy gap of 1.82 eV?
 - A red
 - B orange
 - C green
 - D blue
- 9 The gas in one part of the flame of a burning candle is at a temperature of approximately 1400 K. What band of EM radiation is produced?
 - A X-ray
 - B ultraviolet
 - C visible
 - D infrared

Physics file

When the Apollo astronauts departed from the Moon, they left behind a 50 cm cube corner reflector. Laser pulses from the Earth are aimed at this mirror, and the time they take to return is used to determine the distance between the Earth and the Moon to an accuracy of just a few millimetres.

The invention of the laser in 1960, while it seemed to be just a curiosity at the time, can now be seen as an event of great significance. You have probably witnessed laser light shows at concerts and performances and perhaps seen a laser pointer used at a lecture. More important, however, are the many applications of laser technology in communications and computers. Lasers are used to send signals down optical fibres when you telephone interstate or overseas or use the Internet, computer chips are manufactured using laser technology, CDs and DVDs are read by lasers, and barcode scanners help supermarket queues move faster. Other applications of lasers include eye surgery, industrial cutting and surveying. Lasers are used in computer printers and for measuring the distance to the Moon.

Incoherent and coherent light

In the previous section we looked at the various types of light sources. We saw that there are two main categories: *continuous wide-spectrum light sources* and *narrow-spectrum discrete-line sources*. The radiation emitted by a continuous wide-spectrum light source has a wide range of frequencies across the electromagnetic spectrum. These light sources emit incoherent light: the photons emitted from the source are completely unsynchronised or out of step with each other. The radiation emitted by a narrow-spectrum discrete-line source, on the other hand, has a few specific frequency values. Each element has its own characteristic pattern of spectral lines. Examples are vapour lamps, the light emitted when salts are put into flames and most types of lasers. In vapour lamps and flaming salts, the light is incoherent. Lasers, however, emit *coherent* light.

Coherent light occurs when photons are emitted from a light source in a completely synchronised fashion. This means that the photons are *in phase* with each other as they are emitted from the source.



In **COHERENT LIGHT**, all photons have the same wavelength and are in phase with each other.

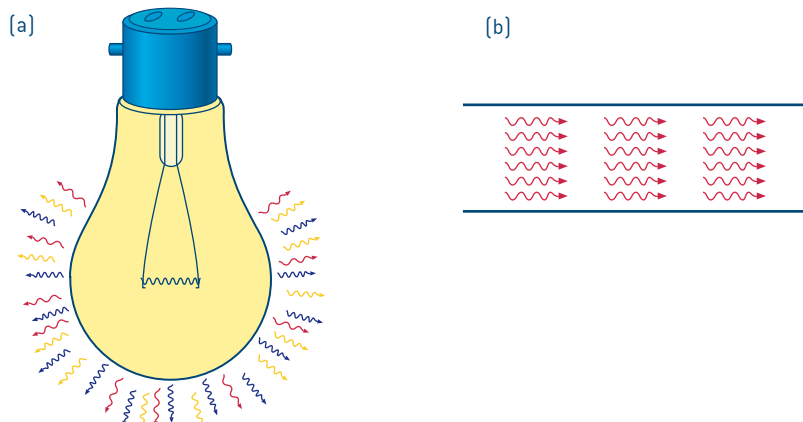


Figure 14.8 (a) A light globe emits photons with a variety of wavelengths and frequencies. Also, the light is incoherent: the photons are out of step with each other. (b) In laser light, the photons have the same wavelength and frequency, i.e. the light is monochromatic. The photons are also in phase, so the laser is a source of coherent light.

The phase synchronisation of laser light makes it useful for observing wave properties of light such as diffraction and interference, which you studied in Unit 4, Area of study 2 'Interactions of light and matter'.

How coherent light is produced

Photons and atoms can interact in three distinct ways:

- 1 Light can be *absorbed* by atoms.
- 2 Atoms can *emit* light in a more or less random fashion.
- 3 Atoms can *emit* a very *ordered* form of light, if stimulated in the right way.

Stimulated absorption

In the first process, a photon is absorbed by an atom. This can only occur if all of the photon's energy is absorbed by a single electron. The energy of the bombarding photon must correspond exactly with the energy difference between energy levels in the atom. When this happens, the photon is completely absorbed and its energy is used to excite the electron to a higher energy level in the atom. In this case:

$$\Delta E = E_2 - E_1 = E_{\text{photon}} = hf = \frac{hc}{\lambda}$$

The *absorption* of energy by the atom is 'stimulated' by the presence of a photon of exactly the right energy (and corresponding wavelength and frequency). The rate at which stimulated absorption occurs is directly proportional to the number of photons of the correct wavelength and the number of atoms in the lower energy state.

Spontaneous emission

An electron in an excited atomic state can spontaneously drop to a more stable energy level, releasing energy in the form of a photon. The photon has a very specific energy (and hence frequency and wavelength), that corresponds to the difference between the initial and final energy states of the electron in the atom. Thus:

$$E_{\text{photon}} = hf = \frac{hc}{\lambda} = \Delta E = E_2 - E_1$$

This process is called *spontaneous emission*. The emission of the photon depends only on the internal properties of the atom and typically occurs about 10^{-8} s after excitation. This time varies, however, so the photons that are emitted via spontaneous emission are not synchronised with each other. Light produced in this way is incoherent. The rate at which spontaneous emission occurs depends directly on the number of electrons that are in the higher energy level.

Stimulated emission

Stimulated absorption and spontaneous emission are the dominant interactions between light and matter. There is also a third process, *stimulated emission*, that only happens under very special conditions.

Stimulated emission happens when an atom in an excited state is irradiated by photons that have an energy equal to the difference between the excited state and a lower-energy state of the atom. In this situation, sometimes

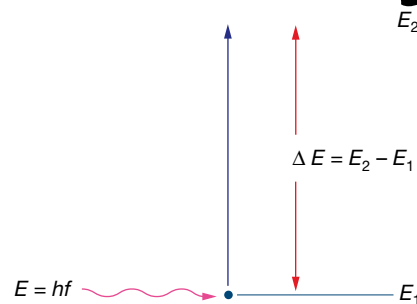


Figure 14.9 For absorption to occur, the energy of the bombarding photon must be equal to the difference between energy levels in the atom. The photon's energy is completely absorbed by the electron it has struck.

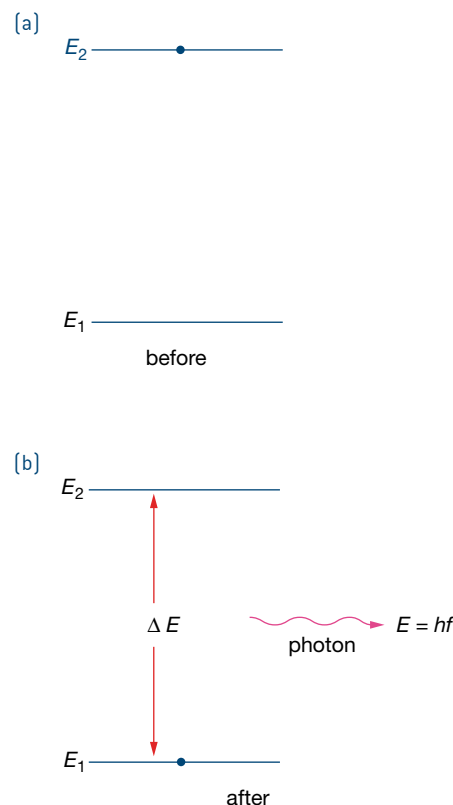


Figure 14.10 (a) This atom is in an excited state. (b) After a very short time, the electron spontaneously drops to a lower energy level and a photon is emitted.

Physics file

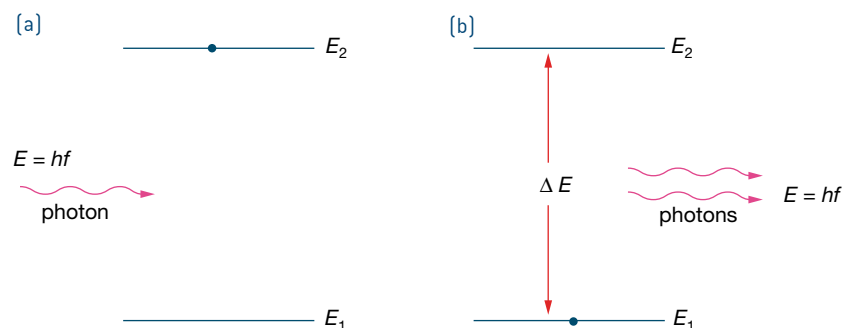
Stimulated emission was theoretically predicted by Albert Einstein in 1913, but it was not actually verified in an experiment until Theodore Maiman produced the first laser light in 1960.

Figure 14.11 (a) An atom in an excited state is irradiated by a photon with energy E equal to the difference in energy levels ΔE . (b) A photon with energy E is emitted. It has the same direction and phase as the first photon.

the incident photon can stimulate the atom so that the excited electron instantaneously loses energy and returns to the lower energy level: as it de-excites, it emits a photon. This second photon has the *same energy* (and wavelength and frequency) as the incident photon. Importantly, the emitted photon is exactly *in phase* with the incident photon and is emitted in the *same direction*. Again:

$$E_{\text{photon}} = hf = \frac{hc}{\lambda} = \Delta E = E_2 - E_1$$

The two photons can now stimulate the same process in other excited atoms, each time releasing photons that are identical and in phase. As this chain reaction continues, a stream of coherent photons is produced.



Physics file

The use of high-output battery-powered laser pointers is restricted. Only use under teacher supervision.

How lasers work

Lasers utilise the principle of stimulated emission of radiation. In fact, laser is an acronym for **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation. The result of the chain reaction described above is a coherent, very intense light beam. This is laser light.

There are a number of conditions necessary for laser light to be produced. First, there must be more atoms in the excited state than in the ground state. This is not the normal condition of matter—it is achieved by ‘pumping’ the lasing material with energy from an external source. This energy is provided by bombarding electrons (electrical pumping) or the bombarding photons (optical pumping).

Second, the excited state that is used must be *metastable*. This means that it must be a relatively stable energy level where the electron takes longer than normal to de-excite and drop to a lower level. In this situation, stimulated emission will occur before spontaneous emission.

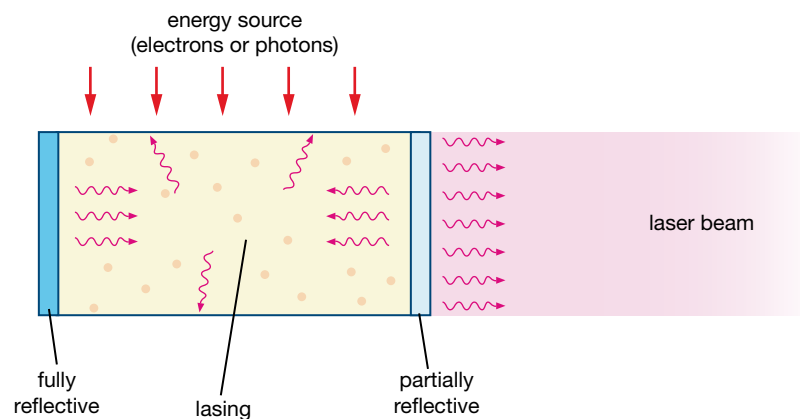


Figure 14.12 As the emitted photons reflect back and forth between the mirrors, further stimulated emissions occur and a laser beam is produced.

Finally, the emitted photons must be used to continue the chain reaction of photon emissions. This is achieved by placing mirrors at the end of the laser tube. One mirror reflects all the photons back, and the other mirror is partially silvered. The mirrors cause the photons to be reflected back and forth through the lasing medium, stimulating further emissions. This is the 'light amplification' part of the process. At the same time, some light escapes through the partially reflective mirror. This is the laser beam.

Types of lasers

There are many different types of lasers and they have many applications.

Gas lasers such as helium, CO₂ and He-Ne lasers use these gases as the lasing medium. Helium and He-Ne lasers produce visible red or green light. These are low-power lasers that are used for scanning bar codes. CO₂ lasers have higher power and produce infrared radiation that is used for cutting through metals and in laser surgery.

Semiconductor lasers, also called diode lasers, are low-power electronic lasers. They are used in CD players, pocket laser pointers and computer printers. Infrared diode lasers are used for hair-removal treatments and for sending signals through long-distance optical fibres.

Excimer lasers use a mixture of reactive gases (such as chlorine and fluorine) and inert gases (such as argon, xenon and krypton) as the lasing medium. They produce a laser beam of ultraviolet radiation. These lasers are used for eye surgery and heart surgery and in the manufacture of micro-machines.

Dye lasers have a liquid containing a coloured dye as the lasing medium. These lasers can be controlled by using prisms and gratings to produce laser light over a wide range of frequencies; that is, they can be tuned by their operator. These tunable lasers are of great benefit when the laser beam is aimed into a material for which the beam colour affects its absorption.

Solid-state lasers use a crystal such as ruby or yttrium aluminium garnet (YAG) doped with neodymium as the lasing material. YAG lasers emit infrared radiation. Some solid-state lasers are also tunable. They are used for eye surgery and high-precision welding.

The wavelengths of some typical lasers are shown in Table 14.2.

Table 14.2 Different types of lasers and their wavelengths

Laser type	Lasing medium	Wavelength (nm)
Gas	carbon dioxide	10 600 (infrared)
Solid state	Nd: YAG	1064 (infrared)
Gas	helium–neon (red)	633 (visible)
Dye	rhodamine 6G (tunable)	570–650 (visible)
Gas	helium–neon (green)	543 (visible)
Gas	argon (blue)	488 (visible)
Excimer	krypton fluoride	248 (ultraviolet)
Excimer	argon fluoride	193 (ultraviolet)

You will have learned about laser diodes in Chapter 5 'Introducing photonics'.



Figure 14.13 Helium lasers are used for scanning bar codes on many products, including groceries, magazines and passports.



Figure 14.14 A pocket laser pointer.



14.2 summary

Coherent light sources: Lasers

- Laser is an acronym for light amplification by stimulated emission of radiation.
- Coherent light is light that is in phase. The photons are synchronised with each other. Laser light is coherent light.
- Atoms that are bombarded with photons will absorb only the photons with energy equal to the difference between their energy levels. The absorption of the photon energy puts the atoms in an excited state.
- Spontaneous emission is the process in which electrons in excited atoms shed their energy by releasing a photon and returning to a lower energy level.
- A laser contains a lasing medium in which most of the atoms are kept in an excited metastable state by 'pumping' the atoms with energy from bombarding electrons or photons.
- When an incident photon of the right energy collides with an excited electron, the electron returns to the ground state and releases a photon that is in phase with the incident photon. This is called stimulated emission. The photons reflect back and forth between two mirrors and collide with other electrons, thus building up a chain reaction of released photons that are all in phase. Some of this coherent light passes through the partially reflective mirror, forming the laser beam.



14.2 questions

Coherent light sources: Lasers

Speed of light in a vacuum $c = 3.00 \times 10^8 \text{ m s}^{-1}$

Planck's constant $h = 6.63 \times 10^{-34} \text{ J s} = 4.14 \times 10^{-15} \text{ eV s}$

- 1 A laser at an observatory is aimed at a reflector on the Moon to help measure the distance from the Earth to the Moon. If the reflected signal takes 2.48 s to return to the observatory, what is the Earth-to-Moon distance (in km) at this point in time?
- 2 A helium–neon laser emits light with a wavelength of 633 nm.
 - a What is the frequency of this light?
 - b Calculate the energy of a single photon.
 - c If the laser had a power rating of 2.5 mW, how many photons would it be emitting each second?
 - d If the laser was aimed at the sky and turned on for just one-tenth of a second, how long (in km) would the resultant laser pulse be?
- 3 Stimulated emission was discussed by Einstein decades before it was actually achieved in the first laser. What is stimulated emission and what is its role in the operation of a laser?
- 4 In a laser, the stimulated emission of photons from the lasing material is transformed into an intense laser beam. Explain how this is achieved.
- 5 In a CO_2 laser, the energy levels between which the excited electrons travel are 0.11 eV apart.
 - a What is the energy of the emitted photons in eV and in joules?
 - b Determine the wavelength (in nm) of the light that is produced by the laser.
 - c Visible light has wavelengths in the range 400–700 nm. Does the CO_2 laser produce visible light?
- 6 An excimer argon–fluoride laser is electrically pumped and produces UV light of wavelength $1.93 \times 10^{-7} \text{ m}$.
 - a Can we see this laser light?
 - b What is used to stimulate the atoms and molecules in the gas mixture?
 - c Calculate the energy of the photons (in eV).
- 7 Identify three differences between light that is emitted from an incandescent light globe and light that is emitted from a He–Ne laser.
- 8 The diagram below shows some of the energy levels for the neon atoms in a He–Ne laser. The gas mixture has been electrically pumped. The electron shown is in the sixth energy level for neon, which is a metastable state. The emitted photons have 2.0 eV of energy.

$n = 6$	_____	20.7 eV
$n = 5$	_____	20.3 eV
$n = 4$	_____	19.8 eV
$n = 3$	_____	18.7 eV
$n = 2$	_____	16.7 eV
$n = 1$	_____	ground state

 - a What is the energy (in eV) of the electrons that were used to pump the gas?
 - b Which energy level does this excited electron drop to?
 - c What is the wavelength of the emitted photons?

14.3 Optical fibres

Photonics is defined as the science of using light to manipulate information and energy. The ancient lighthouse of Alexandria in Egypt is an example that shows how long humans have been using light in some formal system to help them communicate. Nowadays, enormous amounts of data can be sent along a single optical fibre.

Communicating with light

In 1880, William Wheeler patented a series of pipes containing reflectors to channel light through a building. Although the design was not a commercial success, it started some thoughts on the possibilities of using light in this way. Also in 1880, Alexander Graham Bell invented a device called the photophone. Its use involved the transfer of vibrations of a sender's voice to a mirror, where they were translated into variations of light which were passed on to a photocell. The practicalities of relying upon the Sun as a light source and the signal losses due to dispersion hindered the final success of the invention.

The ruby laser, developed in 1960, provided a coherent source for an optical carrier and formed the basis of a breakthrough in optical communications. It was first suggested in the 1960s that information could be transferred by using light in an optical network. Unfortunately, at that time optical fibres had high signal losses of some 1000 dB per kilometre. It is incredible that within a relatively short time, the technology has improved to such an extent that signal losses from optical-fibre cabling are now less than 0.3 dB/km.

The world's first fibre-optic telephone link was operational in the UK in 1977. Today, fibre-optic cable forms the basis of most long-distance telecommunications. Its potential is illustrated by the fact that a single optical fibre can carry 300 000 000 telephone calls at once. By replacing copper cables with optical-fibre cabling, a whole new communication network across the globe through multiple channels of transmission becomes possible.



Capturing light

You will recall from *Heinemann Physics 11* Unit 2, Area of study 2 'Wave-like properties of light', that when light enters a less optically dense medium, it travels faster and is refracted away from the normal. In addition, a weak ray



PRACTICAL ACTIVITY 43

An optical communication system

The photophone is discussed in Chapter 5 'Introducing photonics'.

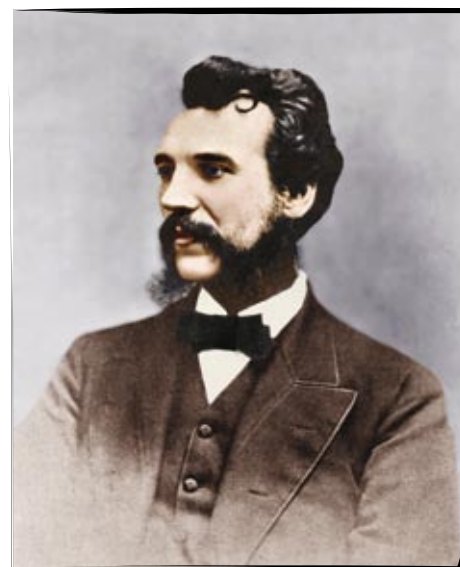


Figure 14.15 Alexander Graham Bell invented the photophone, a device that translated a speaker's voice into variations of light.

Figure 14.16 Optical fibres have led to enormous advances in telecommunications.

You will remember earlier work on total internal reflection from Year 11. Further discussion of the wave nature of light that leads to this behaviour can be found in Chapter 11 'The nature of light'.

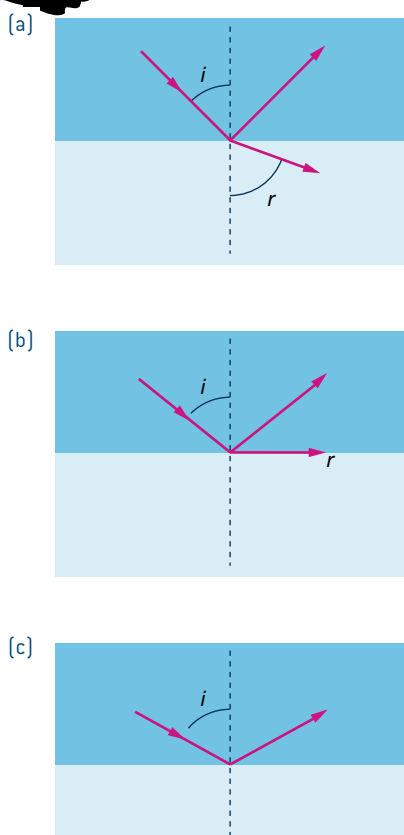


Figure 14.17 A light ray travelling from an optically more dense to a less dense medium. (a) If the angle of incidence is less than the critical angle, the incident ray is refracted away from the normal. (b) At the critical angle, the refracted ray lies along the boundary of the two media. (c) If the angle of incidence is greater than the critical angle, total internal reflection occurs.

Physics file

An endoscope is a flexible light tube that enables doctors to view normally inaccessible parts of the body. It operates by transporting an image by means of total internal reflection along a bundle of optical fibres.

Optical fibres are also discussed in *Heinemann Physics 11* Chapter 8 'Models for light'.

of light is reflected back into the original medium. As the angle of incidence is increased, the angle of refraction also increases, until the refracted ray lies at 90° , along the boundary of the two media. The angle of incidence for which this occurs is called the critical angle, i_c . Any angle of incidence greater than this critical angle will produce no refracted ray at all: instead, all of the incident light will be reflected back inside the optically denser medium. This phenomenon is called *total internal reflection* (TIR).

Recall Snell's law, which states:

$$\frac{\sin i}{\sin r} = \frac{n_2}{n_1}$$

If the incident angle is the critical angle, i_c , at which point $r = 90^\circ$, then

$$\frac{\sin i}{1} = \frac{n_2}{n_1}$$

$$\sin i_c = \frac{n_2}{n_1}$$

An understanding of the principle of TIR is a powerful tool, used in the design of many optical devices. In 1854, the British scientist John Tyndall demonstrated that a light beam could be channelled through a stream of water as a result of total internal reflection, even around a sharp bend. When the water stream begins to break up, the light is no longer trapped within the water and it scatters to our eyes.

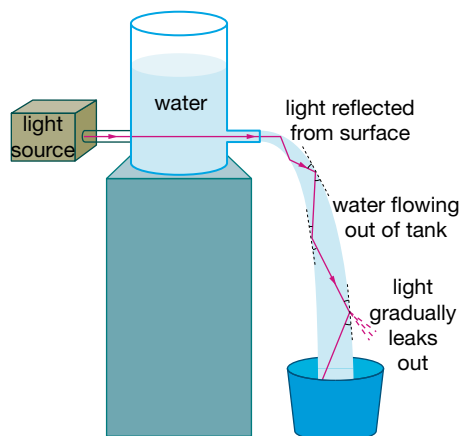


Figure 14.18 Tyndall showed that light could be made to travel around a corner by means of total internal reflection.

Optical fibres as wave guides

Light can be transmitted through a bent glass rod or fibre in the same way that it travelled through Tyndall's water jet. A light beam entering at an angle greater than the critical angle for the medium used will be reflected along a glass fibre to emerge from the other end. If we tape several thousands of glass fibres together, we get a flexible light pipe that can be used to transport much larger quantities of light.

Optical fibres are currently widely used to transmit information in the form of digital pulses of light for such uses as telephone, television and computer communications.

Construction of an optical fibre

Optical fibres are generally made from silica (glass), though plastic fibres may be used for non-communication applications. A standard optical fibre has a diameter of $125 \mu\text{m}$ (0.125 mm). Obviously, this is very narrow! One

micron (μm) is one-millionth of a metre, and a human hair typically has a diameter of about 70 microns.

The presence of dust or grease around the fibre would act to alter the external refractive index, and so the conditions for total internal reflection. As a result, light could be lost through refraction outside the glass core. For this reason, optical fibres are surrounded by a glass *cladding* of slightly lower refractive index than that of the core (i.e. $n_{\text{core}} > n_{\text{cladding}}$). Because the refractive index of the core is only slightly higher than that of the cladding, the critical angle of this interface will be large. This means that the angle of incidence must be small, and only a narrow cone of light will be channelled along the fibre (see Figure 14.19). The silica core of the fibre is usually doped with germanium to increase its refractive index slightly, just above that of the pure silica cladding.

On the outside of the cladding, another layer—made of a sponge-like, shock-absorbing plastic—is used to coat the fibre. This coating is called the *buffer*. It seals the glass surface and protects it from forming small cracks. This structure ensures that the fibre is flexible.

The transmission process

The information travelling through an optical fibre is digital in nature. The electrical information in the signal to be transported—such as speech or pictures—is converted to a digital format (a series of ‘on’ and ‘off’ pulses) in an analogue-to-digital converter (ADC). The resulting signal is impressed onto a light beam produced by a laser or LED in a process called *modulation*. The digital signal is then fed into an optical fibre for transmission.

The radiation used in the transmission process is from the infrared part of the electromagnetic spectrum, just beyond red light in the visible spectrum. Wavelengths of 0.85, 1.3 and 1.5 μm are commonly used, because they lead to a minimal amount of signal loss. Light of these wavelengths suffers less absorption by the glass of the fibre than visible light. Energy losses in transmission are called *attenuation*. We will explore this effect later in this section.

Upon reaching its destination, the signal enters an optical receiver, such as a photodiode, which converts it back into electronic form. The process in which the signal is removed from its carrier wave is called *demodulation*. A decoder processes the signal as required by the end user.

Advantages of optical fibres

The large bandwidth (or range of frequencies) and volume of information transfer demanded by today’s society cannot be catered for using copper cables. An optical-fibre network is more expensive to establish than one based on copper and it has more associated technical difficulties, but the system also has many advantages:

- **high bandwidth.** The capacity of optical fibres is many orders of magnitude greater than that of the best copper cable.
- **low signal attenuation.** The conducting core is pure enough to transmit high-speed signals for 300 km before requiring a repeater, whereas boosters are needed roughly every 50 km for copper media.
- **size.** Optical fibres are lighter, smaller and more flexible than copper cable. A large number of optical fibres can be carried in a cable as thick as a coaxial cable.

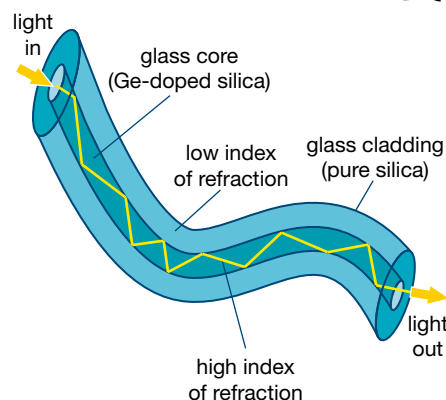


Figure 14.19 Light entering the fibre with a small angle of incidence will have a large critical angle at the core-cladding boundary and will be totally internally reflected along the fibre. A typical optical fibre has a glass core doped with germanium surrounded by a cladding of slightly lower refractive index.

Physics file

Heinrich Lamm transmitted images through a bundle of fibres used as a gastroscope in 1930. Unfortunately, the images were not distinct, because of light losses during transmission. After World War II, Holger Hansen, Abraham van Heel and Harold Hopkins each developed fibre bundles to transmit images. Hansen coated his fibre with margarine to reduce light loss through refraction into the surrounding air. An American physicist, Brian O’Brien, suggested the use of some form of cladding to van Heel. Van Heel then used plastic and beeswax to coat the fibres. Great advances in the reduction of light losses occurred with the development of the monochromatic light of the laser in the 1960s and with the purification of the glass used to make the fibres in the 1970s.



PRACTICAL ACTIVITY 44

Total internal reflection in fibres and light pipes

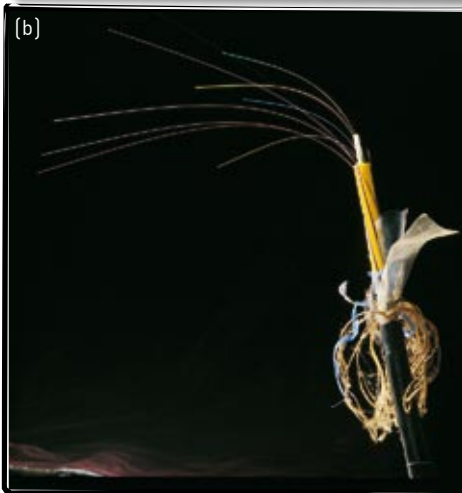
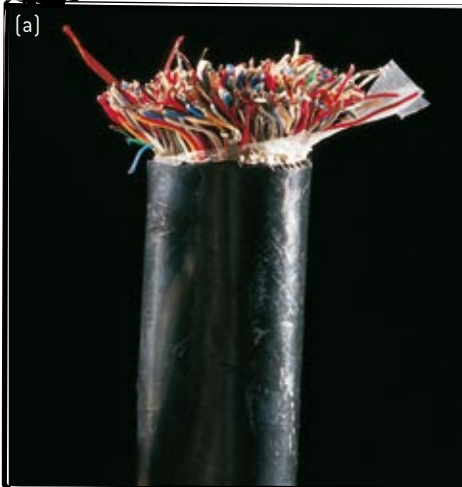


Figure 14.20 Optical fibres versus copper cables: 600 twisted-pair copper cable (a) carries 600 conversations. Six coaxial copper cables carry 2700 conversations. Now, one single optical fibre (b) can carry millions of conversations.

- *electrical isolation.* Optical fibres do not suffer from 'crosstalk' or radio interference.
- *security.* It is very difficult to 'tap' into an optical fibre. In order for this to occur, the fibre must be broken, in which case the tap will be detectable as no signal is detected at the other end.

Types of optical fibres

As mentioned, an optical fibre is a wave guide. Once inside the fibre, any ray incident on the walls at an angle greater than the critical angle will be reflected along the inside of the fibre. Some rays will undergo more reflections than others, depending on their incident angle when entering the fibre. Some will undergo destructive interference and will die out. Rays that undergo constructive interference will continue along the length of the cable. In this way, only rays of particular pathways will be guided through the fibre. These sets of light waves that travel in allowed directions of propagation are called *modes*.

Higher-order modes undergo more reflection than lower-order modes, and as a result they actually travel further. This creates a problem: higher-order modes will take longer to reach the end of the end of the fibre than lower-order modes. Different types of cable have been designed to help minimise this problem. There are three basic types of optical fibre: step-index fibre, graded-index multimode fibre and single-mode fibre.

Step-index fibre

The step-index fibre is a multimode fibre, which means it will support more than one propagating mode. It usually has a core diameter of $50\text{ }\mu\text{m}$, which is large enough to allow the cable to easily be coupled to light sources. The refractive index is constant in value throughout the glass core and then abruptly decreases at the core-cladding interface. All light rays inside the core travel at the same velocity, due to the constant refractive index in this region.

Because the higher-order modes take longer to travel the length of the fibre, any sharp pulse of light entering this fibre will become more spread out as it travels further. This effect limits how close the pulses can be before

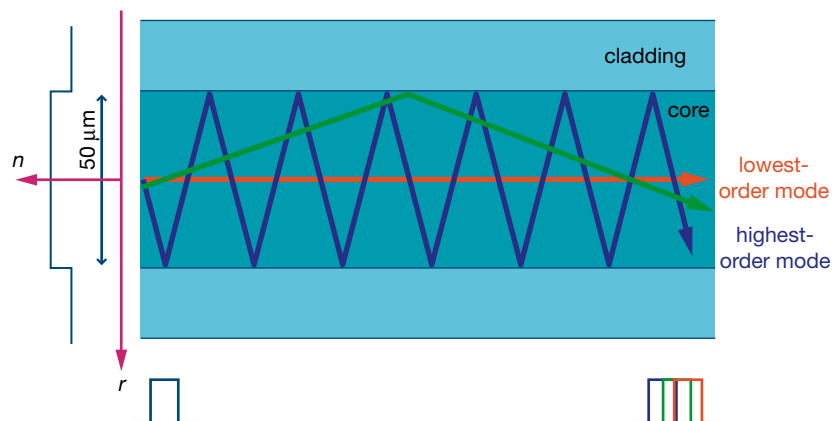


Figure 14.21 Different modes will arrive at the end of a step-index fibre at different times. This leads to the spreading of a narrow input signal by the time it reaches the other end of the fibre.

they overlap and become indistinguishable from each other. The spreading out—or loss of definition—of the pulses is referred to as *modal dispersion*.

In this simple type of fibre, modal dispersion can amount to about 100 ns (or 10^{-7} s) per kilometre: a very sharp pulse of light would be spread over 100 ns after a kilometre of travel. This means that less than 10^7 pulses can be sent along a kilometre of this cable each second. That may seem a lot, but it is less than 10 MHz, which in today's terms corresponds to a fairly slow rate of information transfer: compare this figure with the speed of a typical home computer, which would easily be 1000 MHz.

Graded-index multimode fibre

The graded-index multimode fibre, like the step-index fibre, carries a number of modes of propagation and has a diameter of 50 μm . However, the refractive index is high in value at the core of the fibre and gradually decreases in value extending outwards to the core-cladding boundary. Light travelling through the centre of the core has the slowest velocity, because the refractive index is high in this region.

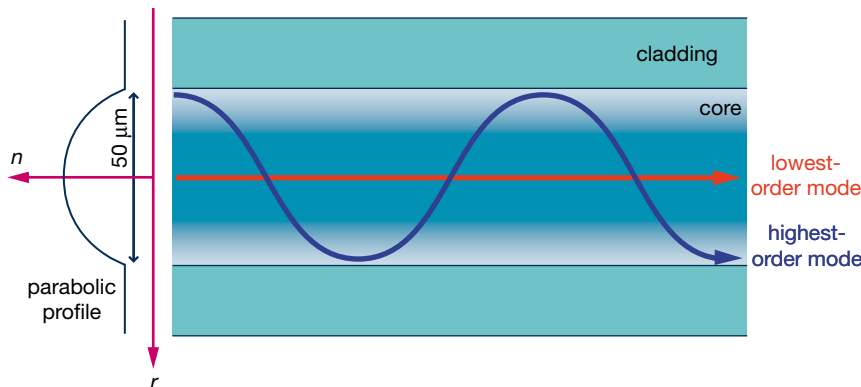


Figure 14.22 The graded refractive index in this type of multimode fibre results in the lowest-order mode travelling the shortest distance at the slowest velocity. Higher-order modes travel the greatest distance at a faster velocity. Hence, all modes transit in approximately the same time.

This has the effect of slowing the travel time of lower-order modes. Higher-order modes travel greater distances, but with increased velocity due to the lower refractive index towards the edge. The profile of the core's refractive index is parabolic, which ensures that all modes transit the fibre in approximately the same time. As a result, modal dispersion is reduced to about 1 ns per kilometre of signal transmission.

Single-mode fibre

Single-mode fibre has a core diameter of only 5 μm . It is so small that only the lowest-order or fundamental mode can propagate. The core refractive index is constant, with an abrupt decrease at the core-cladding interface. Because only a single mode travels through the fibre, there is zero modal dispersion.

Clearly, single-mode fibres are advantageous to use because they suffer no modal dispersion. Unfortunately they are more expensive to make, and coupling light into and out of such a small diameter core is very difficult. Single-mode fibres are used increasingly for long-haul transmission systems, but multimode fibres are satisfactory for many short distance and low-bit-rate applications.

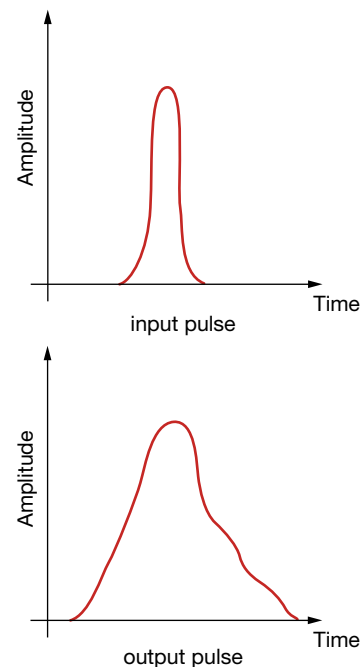


Figure 14.23 Modal dispersion causes the spreading of a narrow input pulse, because different modes arrive at the exit of the fibre in different times.



Figure 14.25 A 1.25 GByte/s multimode fibre with the termination removed to show the internal structure.

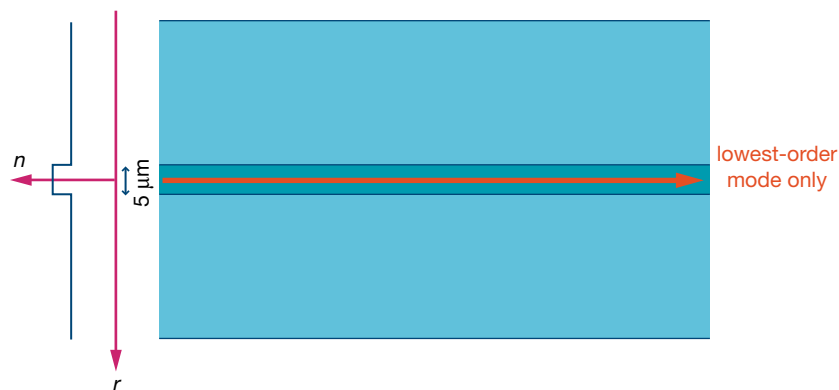


Figure 14.24 Only a single mode is possible in this fibre, because the diameter of the core is of the same order of magnitude as the wavelength of light.

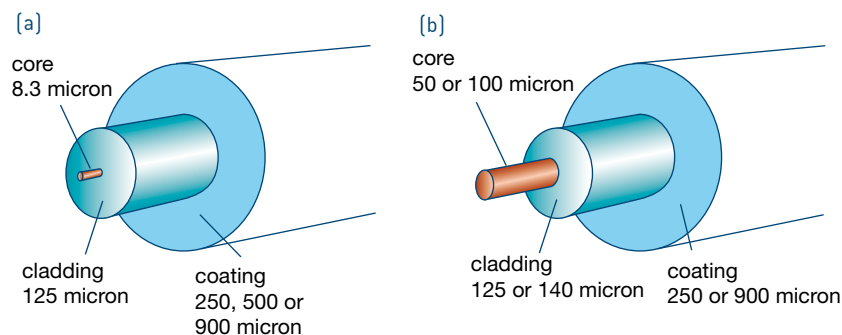


Figure 14.26 (a) Single-mode fibres, such as those used in high-speed telecommunications, result in much less distortion of the optical system. (b) Multimode fibres, such as those used in local area networks, are cheaper to produce but have more distortion problems. They are adequate for use over shorter distances.

Obviously, a communication link does not consist of one single optical fibre. Typically, between 2 and 36 fibres are grouped together inside a tube with a gel-like substance between them. Each tube is called an *optical unit*. A cable is formed by placing optical units around a strength member, usually made from braided Kevlar, which bears the tension in the cable so that the fibres are not stressed to a large degree. The entire cable is then covered in a PVC jacket to protect it from its surroundings.

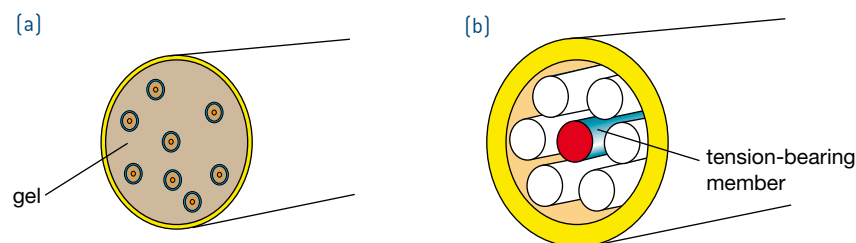


Figure 14.27 (a) An optical unit typically houses between 2 and 36 fibres. The number of fibres is always even, because each pair of fibres carries light in opposite directions. (b) An optical cable is formed by placing optical units around a tension-bearing member. The strength of this member will depend upon the distance the cable must span. The entire cable is then covered with a plastic coating to protect it from the weather.

Worked example 14.3A

A step-index multimode fibre has a core of Ge-doped silica glass, of refractive index 1.500. The pure-glass cladding has a refractive index of 1.480. Calculate the critical angle of the core-cladding boundary of this optical fibre.

Solution

$$\begin{aligned}n_{\text{core}} &= 1.500, n_{\text{cladding}} = 1.480 \\ \sin i_c &= \frac{n_2}{n_1} \\ i_c &= \sin^{-1} \frac{1.480}{1.500} \\ &= 80.63^\circ\end{aligned}$$

Trapping light

As you know, optical fibres work because they channel light within themselves and move it to another place. This process will only be effective if we can first trap light inside the fibre, and then minimise the losses as the light is reflected along the fibre. We will now consider the characteristics of a fibre concerned with trapping light inside it.

Acceptance angle

What determines the light-gathering ability of an optical fibre? If a ray of light directed towards a fibre is incident at an angle less than the critical angle, it will be refracted into the cladding instead of undergoing TIR and continuing down the fibre. The maximum angle at which a ray can enter the fibre and then propagate by TIR is called the *acceptance angle*, θ_1 .



The **ACCEPTANCE ANGLE**, θ_1 , is the maximum angle at which a ray can enter an optical fibre and then propagate by total internal reflection.

Due to the cylindrical geometry of a fibre (about its central axis), the acceptance angle actually corresponds to a cone of rays that may enter the fibre, as can be seen in Figure 14.28. This *cone of acceptance* has an apex angle of $2\theta_1$. Any ray that is incident upon the fibre within this cone will be totally internally reflected through the core of the fibre.

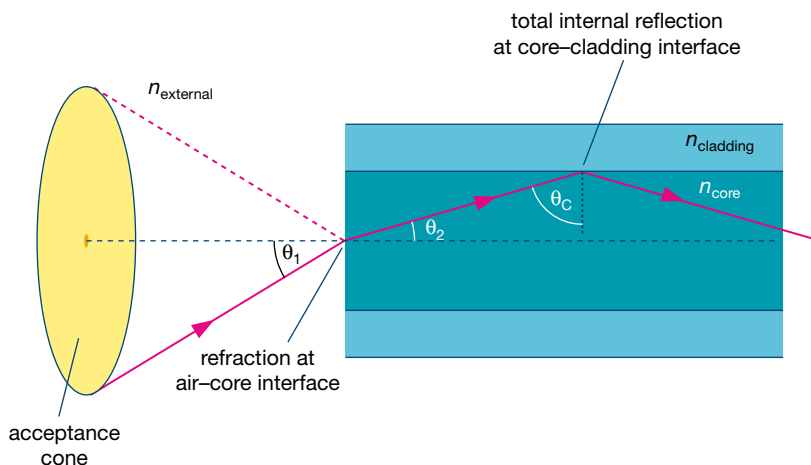
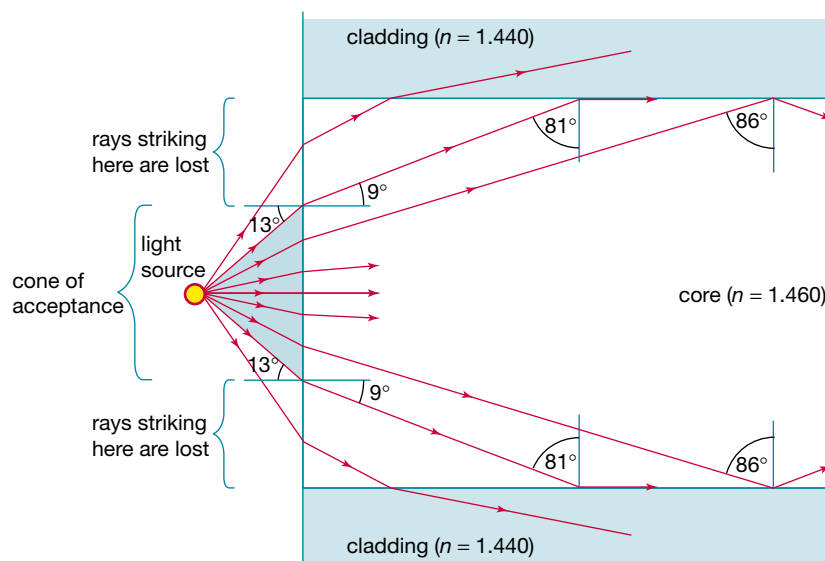


Figure 14.28 θ_c is the critical angle inside the core. For TIR to occur, θ_2 (equal to $90^\circ - \theta_c$) is the largest refracted angle at which light can enter the core. θ_1 is referred to as the acceptance angle. The acceptance cone is the range of incident angles which will result in light undergoing TIR in the core of the fibre. Rays outside this cone will be lost in the cladding. The acceptance angle depends on the refractive indices of the core, the cladding and the external medium.

Figure 14.29 This diagram shows the cone of acceptance of rays that will be totally internally reflected through a glass core of refractive index 1.460 with a cladding of refractive index 1.440. The critical angle for the core–cladding interface is 81° . Outer rays from the light source that enter the core at angles of incidence greater than 13° will continue to strike the core–cladding boundary at an angle of incidence less than 81° . These rays will therefore be transmitted into the cladding and be lost. Rays originally within the cone of acceptance strike the cladding at an angle greater than 81° and are therefore totally internally reflected. These reflected rays carry the signal along the fibre.



Numerical aperture

Snell's law tells us that

$$n_{\text{ext}} \sin \theta_1 = n_{\text{core}} \sin \theta_2$$

where n_{ext} and n_{core} are the refractive indices of the external medium and the core, respectively. For a given core and cladding, $n_{\text{ext}} \sin \theta_1$ has a certain value which cannot be exceeded if light is to travel down the fibre. This value ($n_{\text{ext}} \sin \theta_1$) is known as the *numerical aperture* or NA.



The NUMERICAL APERTURE of an optical fibre is a measure of how readily it will capture light. It has a value between zero and one.

$$NA = n_{\text{ext}} \sin \theta_1$$

As we have seen, this quantity depends upon the critical angle inside the fibre, which in turn depends upon the difference between the refractive indices of the core and cladding. A little algebra and trigonometry enable us to find that the numerical aperture is actually equal to the square root of the difference of the squares of the two refractive indices:

$$NA = n_{\text{ext}} \sin \theta_1 = \sqrt{n_{\text{core}}^2 - n_{\text{cladding}}^2}$$

where n_{ext} = refractive index outside fibre

θ_1 = acceptance angle

n_{core} = refractive index of the core

n_{cladding} = refractive index of the cladding.

The larger the difference in refractive indices, the more light can get into the fibre. The numerical aperture lies between zero and one. A numerical aperture of 0 corresponds to the fibre gathering no light, because there is no TIR: $\theta_1 = 0^\circ$. A numerical aperture of 1 would correspond to a fibre that gathers all light incident upon it: $\theta_1 = 90^\circ$, meaning that all incident light would undergo TIR in the core. A typical multimode fibre has a numerical aperture of between 0.2 and 0.3.

If the fibre is located in air, then $n_{\text{ext}} = 1$ and so:

$$NA = \sin \theta_1$$

If the fibre is immersed in water, then the external refractive index has a value greater than one. In this case, the numerical aperture is unchanged, but the acceptance angle, θ_1 , is smaller.

Worked example 14.3B

The step-index optical fibre in Worked example 14.3A had a doped-glass core of refractive index 1.500 and a surrounding cladding of refractive index 1.480. Calculate:

- a** the numerical aperture for this fibre
- b** the acceptance angle for this fibre
 - i** if the fibre is in air ($n = 1.000$)
 - ii** if the fibre is immersed in water ($n = 1.330$).

Solution

- a** $NA = n_{\text{ext}} \sin \theta_1$
 $= \sqrt{n_{\text{core}}^2 - n_{\text{cladding}}^2}$
 $= \sqrt{1.500^2 - 1.480^2}$
 $= 0.244$
- b i** In air: $NA = n_{\text{ext}} \sin \theta_1$
 $n_{\text{ext}} \sin \theta_1 = 0.244$
 $\sin \theta_1 = 0.244$
 $\theta_1 = 14.1^\circ$
- ii** In water: $NA = n_{\text{ext}} \sin \theta_1$
 $n_{\text{ext}} \sin \theta_1 = 0.244$
 $\sin \theta_1 = \frac{0.244}{1.330}$
 $\theta_1 = 10.6^\circ$

The acceptance angle of the fibre is 14.1° in air, but only 10.6° in water.

Losing light

In any communication system, some of the signal will be lost as it travels from its source to its destination. *Attenuation* is the loss of signal strength (in the case of optical fibres, optical power) along the length of the cable. This loss is measured in decibels (dB), which means that it's not determined by a simple fraction but follows a logarithmic scale: the loss of a certain fraction of power corresponds to the subtraction of a certain numerical value. For example, a loss of half the signal power corresponds to an attenuation of 3 dB, while if only one-quarter of the original power arrives at the destination, the attenuation is 6 dB.

The attenuation in decibels is defined by the following equation:

$$\Delta P = -10 \log \frac{P_{\text{in}}}{P_{\text{out}}}$$

where ΔP = attenuation = power loss (dB)

P_{in} = power in (W)

P_{out} = power out (W).

There are a number of sources of signal attenuation. The most significant ones are Rayleigh scattering, absorption by impurities, bending losses and Fresnel-reflection loss.

Physics file

Typical values of attenuation range from 10 dB km^{-1} for a step-index multimode fibre with a 850 nm light source to less than 0.2 dB km^{-1} for a single-mode fibre with a 1550 nm light source.

Table 14.3 Some values of attenuation with corresponding percentages of power loss

Attenuation [dB]	Power loss [%]
20	99
10	90
3	50
1	50
0.1	2

Physics file

Signal attenuation is measured by an internationally recognised reference test method. It is called the 'cut-back technique'. Light of a known spectral make-up is directed into a fibre and the power transmitted is recorded as a function of wavelength. The fibre is then cut back by a few metres and the process is repeated. The comparison of the power transmitted from the two tests is used to calculate the spectral attenuation.

Rayleigh scattering

During the manufacturing process of the optical fibre, non-uniform mixing of the glass produces very tiny variations in density as it is drawn into a fibre and cools. Even though the affected areas can be smaller than the wavelength of the light, they scatter a small amount of light, which is either lost from the signal or degrades it. This type of scattering is called *Rayleigh scattering*. The degree to which it occurs depends upon the relative size of the irregularities and the wavelength of light involved. The larger the irregularities are, the greater the scattering.

The amount of Rayleigh scattering is inversely proportional to the fourth power of the wavelength. This means that halving the wavelength will increase the scattering by a factor of 16. As a result, it is advantageous to use light of longer wavelengths in fibre optics, in order to minimise this scattering of the signal. For example, it is better to use light with a wavelength of 1550 nm than of 850 nm.

Absorption by impurities

As light travels along a fibre, impurities in the glass absorb different components from the beam. Captured energy is converted to heat and is lost from the signal. The most significant breakthrough in achieving better attenuation levels has come about through the further purification of the glass used to make the fibres.

Overall attenuation varies with wavelength. An understanding of this correlation is important in order to minimise signal losses by selecting the appropriate wavelengths for particular fibres. Figure 14.30 illustrates the

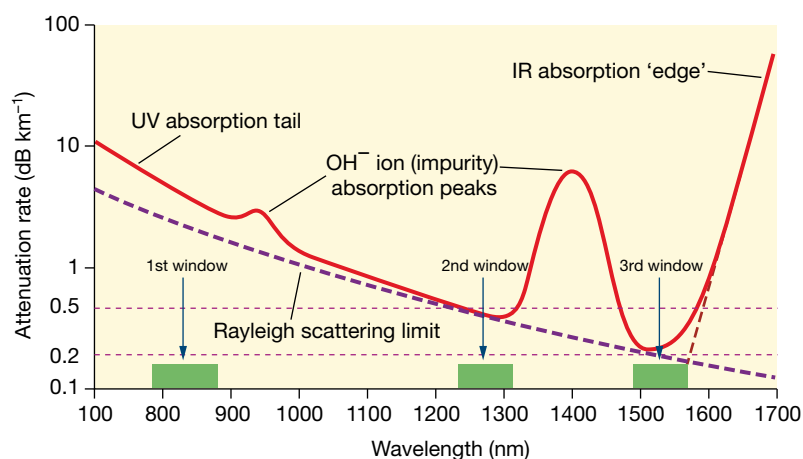


Figure 14.30 Silica-glass fibres have their minimum attenuation (0.2 dB km^{-1}) for wavelengths of 1550 nm. The first, second and third telecommunications 'window' for minimum attenuation are shown at around 850, 1300 and 1550 nm.

way in which wavelength affects signal attenuation in fibres. Two peaks in attenuation can be observed, which correspond to the resonances from O–H bonds which occur even with ultra-low concentrations of water contamination. Generally, the level of attenuation can be seen to decrease with increasing wavelength, due to the lesser Rayleigh scattering, up until about 1.5 μm . The useful minima for signal losses are found around wavelengths of 1.3 μm (1300 nm) and 1.55 μm (1550 nm). As a result, these wavelengths are most commonly employed in optical-fibre systems. Note the rapid increase in attenuation in the infrared region, due to absorption by the Si–O and Ge–O bonds.

Bending losses

The term *macro bending* describes the situation in which a bend in a fibre causes light to hit the core–cladding interface at an angle less than its critical angle. As a result, it is refracted through the cladding and lost from the fibre.

Light is also lost through a process called *micro bending*. Microbends are tiny kinks in the fibre, within the plastic jacket. As the fibre cools during the manufacturing process, the jacket may shrink more than the fibre and thus kink the fibre. Alternatively, these tiny variations can form when the fibre is placed under stress. Their presence can cause light to be incident at angles less than the critical angle, as with macro bending, and similarly be lost as a result. This process can be put to good use in the application of sensors, which will be explored in the next section.

Fresnel reflection

An optical-fibre cable can contain many splices, joining separate sections together. At the site of each connector or splice, some light will be reflected back from the end face of the fibre. This process is called *Fresnel reflection*. The reflection increases with the angle of incidence. As for lower-order modes in a fibre, this angle is close to zero, the loss due to Fresnel reflection is minimal for these modes at around 4%.

The proportion of power reflected at the end face of a fibre is given—for rays normal to the face—by the equation:

$$I_R = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2$$

where I_R = fraction of incident power reflected

n_1 = refractive index of the core

n_2 = refractive index external to the fibre end.

So, for example, in the case of a fibre core of refractive index 1.50 (n_1) positioned in air of refractive index 1.00 (n_2), $I_R = (0.50/2.50)^2 = 0.04$. This means that 4% of the power transmitted along this fibre will be lost due to Fresnel reflection.

Because the amount of reflection increases with the difference between the refractive indices inside and outside the core, the use of gels or oils with a refractive index similar to that of the core will reduce the amount of Fresnel reflection.

PRACTICAL ACTIVITY 45

Fibre optic cladding

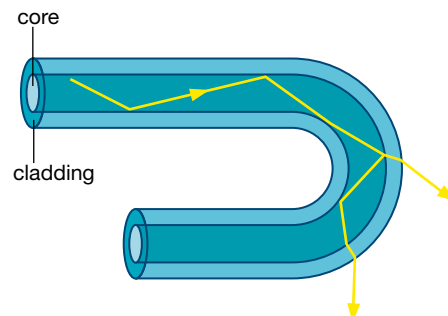


Figure 14.31 A sharp bend in an optical fibre can result in light rays being refracted into the cladding and lost from the signal. These are called 'leaky modes'.

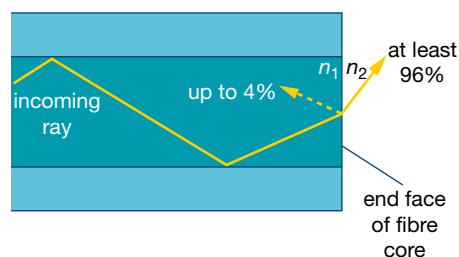


Figure 14.32 Fresnel reflection refers to the small fraction of light that is reflected back at the end of a fibre.

Physics file

Pulse spreading affects the bandwidth and distance limit of an optical fibre. This effect is described by the bandwidth–distance product (BDP), which relates to the fibre's bandwidth multiplied by its length. It has a constant value for a particular fibre.

Usually the bandwidth of a fibre over a 1 km distance is stated. For example, for a fibre with a BDP of 20 MHz km:

- 1 km of the fibre would have a bandwidth of 20 MHz
- 5 km would have a bandwidth of 4 MHz
- 10 km would have a bandwidth of 2 MHz
- 20 km would have a bandwidth of 1 MHz.

Typical BDPs are:

- single-mode fibre: 500–1500 MHz km
- graded-index fibre: 100–1000 MHz km
- step-index fibre: 6–25 MHz km.



PRACTICAL ACTIVITY 46

Wavelength of LEDs

Physics file

We can battle the attenuation and dispersion effects of optical-fibre signals using optical repeaters. Optical repeaters are optical devices that combat attenuation by amplifying the signal. Because of this amplification, spans of fibre cable over some 80 km stretches are now possible.

Regenerators consist of electronic and optical components. First they convert the light signal into an electronic signal. They then retune and reshape this signal to remove the effects of dispersion and noise. The clean signal is converted back into a light signal, which is optically amplified and sent through the system. It is imagined that the optical-fibre networks of the future will contain only optical components and have no electrical parts. Such systems will be much faster, as the light signal will not have to be converted into an electronic signal and vice versa.

Material dispersion and bandwidth

The data relayed along an optical fibre is transmitted in the form of pulses of light energy following each other in quick succession. We have considered the phenomenon of modal dispersion, in which different modes travel different path lengths through the optical fibre, leading to the dispersion or spreading of their arrival over time. The spreading of the signal over time is called 'pulse spreading', and was illustrated in Figure 14.23.

In addition to modal dispersion, another type of dispersion also causes pulse spreading. It is called *material* or *chromatic dispersion*. You will recall that white light separates into various colours when passing through a prism. This occurs because the refractive index of a material actually varies with the frequency (or wavelength) of the incident light. As a result, light rays of different frequencies travel through a particular material at slightly different speeds. Because the light used in any fibre-optic communication system has a range of wavelength components, these will travel at slightly differing speeds. Hence, material dispersion occurs and any narrow pulse is spread, even in the core of a single-mode fibre.

The *total dispersion* of a signal is the sum of the material and modal dispersion effects. These dispersion effects have implications for the *bandwidth*. The bandwidth refers to the maximum frequency of signal that can be sent along a fibre. If the frequency of the signal is too high, then the pulses will overlap at the other end of the fibre and the information will be meaningless for the user. In combination with the limitations placed on the bandwidth, there are limitations on the distance a signal may be transmitted. Increasing the distance will increase the pulse spreading.

Optical transmitters

Two types of light sources are used with optical fibres: LEDs and laser diodes.

You have already explored the nature of LEDs. They can emit infrared radiation at wavelengths commonly used in fibres: 850, 1300 and 1550 nm. LEDs have long lifetimes and are cheap. However, they emit light over a relatively large range of angles, which leads to light losses in coupling with the fibre. Also, LEDs do not emit light of a pure frequency. Their *spectral spread* is about 30–40 nm. This leads to a degree of material dispersion, and hence pulse spreading, in an optical fibre. As a result, LEDs tend to only be used in multimode fibres and in low-bit-rate digital systems where pulse spreading will not make an impact.

Laser diodes can be made to emit light at 1300 nm or 1550 nm for use in optical fibres. Despite having a shorter lifetime and being more expensive to



Figure 14.33 A digital signal before and after regeneration.

produce than LEDs, laser diodes have the great advantage that the spectral spread in the wavelengths of the light they transmit is only 2 nm or less. Thus, much less pulse spreading of the signal occurs. Laser diodes can also emit light over a very small area, which means that the angle of incidence for light entering the fibre can be accurately controlled. Lasers diodes are used for high-bit-rate fibre systems with single-mode fibres.

See Physics in action 'Light—an electromagnetic wave', page 390.

Modal patterns

A *mode* is a particular set of light waves that travel through an optical fibre. Light is a transverse electromagnetic wave, as you will recall from Chapter 9. It consists of oscillating electric and magnetic fields at right angles to the direction of propagation. What implications does this have for the way light travels through an optical fibre?

As the waves travel down the fibre, interference patterns are set up between different wavefronts reflecting off the sides of the core. Some of these will interfere constructively, others destructively. Only modes that involve constructive interference between wavefronts will propagate along the fibre. This explains why only a certain number of modes can propagate along the fibre. One mode is always allowed through a fibre: the lowest-order, or fundamental, mode. It is the set of waves that travels directly through the centre of the fibre.

A calculation of the number of modes through a fibre depends upon finding solutions to Maxwell's equations (which describe electromagnetic waves) applied to a cylinder. The reflection of waves back and forth across the fibre results in a sort of standing-wave pattern across the fibre (not along it!). The number of modes allowed along a fibre is proportional to the radius of the fibre core and the numerical aperture and is inversely proportional to the wavelength of the light being used. In other words, the larger the ratio of the core diameter to the wavelength, the larger the number of possible modes. Some possible modes are shown in Figure 14.34.

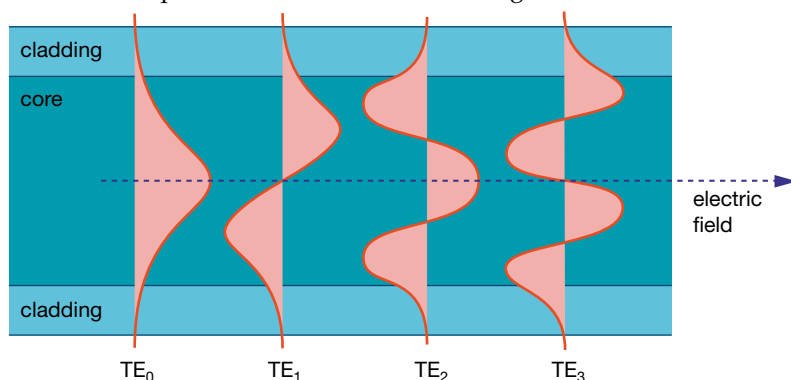


Figure 14.34 The electric and magnetic fields of light waves propagate at right angles to each other and are transverse as they propagate along the axis of the fibre. The field patterns shown here are 'transverse electric' [TE]. This diagram shows the first four TE mode field patterns. TE_0 has one field maximum at the centre of the glass core: this is the fundamental mode. Some of the modes with higher numbers of maxima actually penetrate the cladding. This results in an exchange of power between modes, leading to power losses from the core modes.

When looking at the patterns formed at the exit end of an optical fibre—such as in Figure 14.35—we can see bright and dark regions, created as a result of the constructive and destructive interference associated with the wavefronts. These patterns are symmetrical about the centre of the beam.

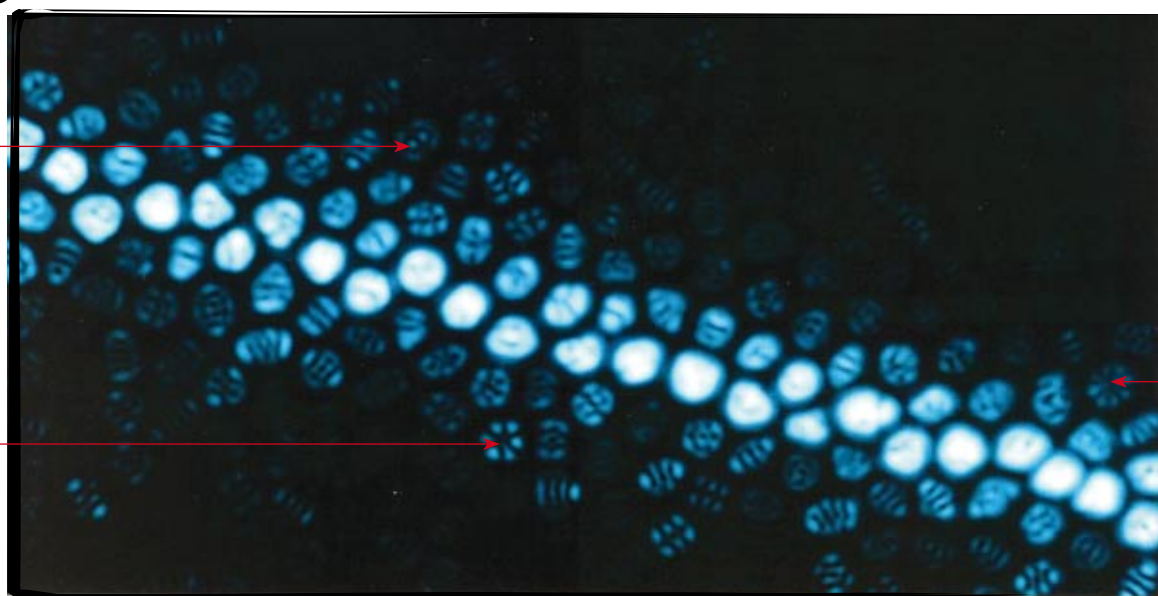


Figure 14.35 These light and dark patterns show the nodes due to the wave nature of light as it propagates through optical fibres.

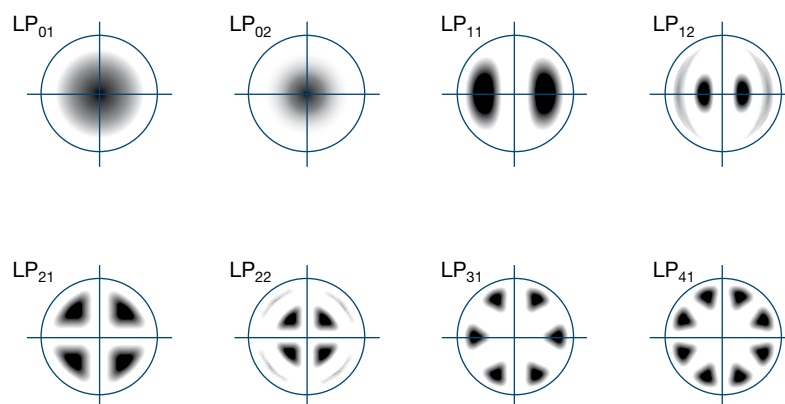


Figure 14.36 A range of linearly polarised (LP) modal patterns of particular modes of propagation along an optical fibre.



14.3 summary

Optical fibres

- Optical fibres are used as wave guides to channel light by means of total internal reflection (TIR). They are made of glass or plastic and are surrounded by a cladding which has a slightly lower refractive index than the core.
- Optical fibres offer advantages over copper cabling networks. They have a much higher bandwidth, are lighter, smaller and more flexible than copper cable, are isolated from interference and offer greater security.
- Only rays of particular pathways will be guided through a fibre. The sets of light waves that travel in allowed directions of propagation are called modes.
- Optical fibres are available in three basic types:
 - step-index multimode fibre*, with a constant refractive index in the core and an abrupt discontinuity at the core-cladding boundary
 - graded-index multimode fibre*, with a gradually decreasing refractive index radiating outwards from the centre of the core
 - single-mode fibre*, with a constant refractive index in the core, which is so narrow that it only allows propagation of the fundamental mode.
- The maximum angle at which a ray can enter the fibre and then propagate by total internal reflection is called the acceptance angle, θ_1 .



- The numerical aperture is a measure of a fibre's light-gathering ability. It is defined as:

$$NA = n_{\text{ext}} \sin \theta_1$$

and is equal to: $\sqrt{n_{\text{core}}^2 - n_{\text{cladding}}^2}$

- Attenuation* is the loss of signal strength, or optical power, measured from the source to the destination. This loss is measured in decibels (dB):

$$\Delta P \text{ (dB)} = -10 \log \frac{P_{\text{in}}}{P_{\text{out}}}$$

- Attenuation results from Rayleigh scattering, absorption by impurities, bending losses and Fresnel-reflection loss.

- Infrared wavelengths of 1.3 μm (1300 nm) and 1.55 μm (1550 nm) correspond to minimum attenuation levels in silica doped with germanium. These are the wavelengths most commonly employed in optical-fibre systems.

- The bandwidth of an optical fibre is limited by:
 - material dispersion*: the spreading of a pulse due to the differing speeds of the range of wavelengths making up the incident light
 - modal dispersion*: the spreading of a pulse due to different modes travelling different distances to reach the end of a fibre.



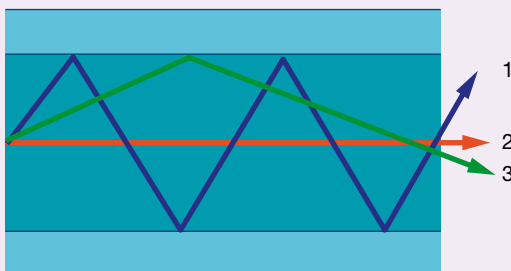
14.3 questions

Optical fibres

- List and explain five advantages of a network of optical fibres compared with copper cabling.
- A step-index multimode fibre has a core:
 - of constant refractive index and a cladding of constant but lower refractive index
 - of constant refractive index and a cladding of constant but higher refractive index
 - with a refractive index that gradually decreases in value travelling outwards from its centre and a cladding of constant refractive index
 - with a refractive index that gradually increases in value travelling outwards from its centre and a cladding of constant refractive index.
- What is the typical core diameter of a step-index multimode fibre?

The following information relates to questions 4–6.

The diagram shows a number of modes propagating along a step-index multimode fibre.



- Which number represents:
 - the fundamental mode?
 - the highest-order mode?

- Given that the velocity is constant for each mode, in which order will the modes arrive at the exit end of the fibre?
- The spacing of the arrival of different modes is called *modal dispersion*. Given that the light source used in this instance is an LED, name and explain a second process that will produce further spreading of the input pulse.

The following information relates to questions 7–9.

An optical fibre in air has a glass core and cladding with refractive indices of 1.458 and 1.440 respectively.

- Calculate the numerical aperture of the fibre.
- Calculate the acceptance angle for which incident light will be channelled.
- What would be the effect on the NA and the acceptance angle if the fibre were immersed in gel of refractive index 1.26?
- Making reference to Figure 14.30, describe how signal attenuation varies with wavelength for an optical fibre.
- A mean power of 1.60 mW is launched into a one-kilometre length of optical fibre with an attenuation loss of 0.30 dB. What is the output power received from the fibre?

14.4 Applications of optical fibres

Optical fibres find more and more applications in modern society. One important area in which they are used is telecommunications. The Internet could not have developed the way it has without optical fibres. Advances in the production of optical fibres and components have also made the development possible of a wide variety of fibre-optic sensors and imaging systems.

Telecommunications

The world we live in has demands of bandwidth and speed that cannot be met by a solely copper-based communication network. As a result, a network based on optical fibres is rapidly becoming a reality. A single-mode optical-fibre cable has a bandwidth of some 10 000 MHz, compared to around 100 MHz for coaxial cable. A single optical fibre is capable of transporting one million conversations! Network managers are looking towards this technology to meet future requirements. In the 1980s, more than 3 million kilometres of fibre-optic cabling were installed in the United States, replacing enormous stretches of copper cabling for long-distance communications. In Australia, long-distance optical fibres cover over 1.5 million kilometres.



Figure 14.37 Copper cables are being replaced with optical fibres for long-distance telephone communications.

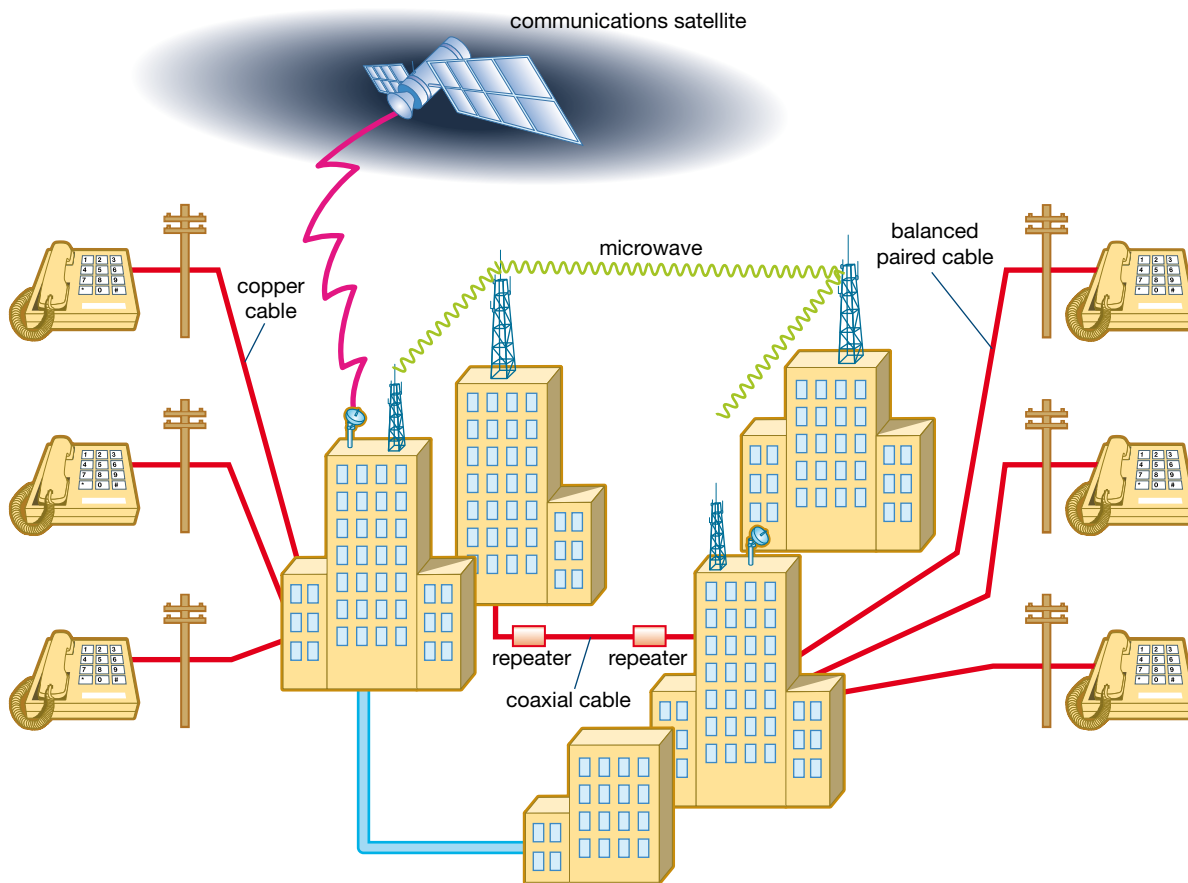



Figure 14.38 A number of communication mechanisms are shown here: copper or coaxial cables, microwave surface links, microwave satellite links and optical fibres.



Replacing copper cables with optical fibres has huge implications, not only for signal bandwidth, but also for the size and weight of the cabling. A bundle of copper wire with a diameter of 76 mm can be replaced with a 6 mm optical-fibre cable; some 3.5 kg of optical fibre is as effective as almost 100 kg of copper cabling. Thus, optical-fibre cabling is particularly beneficial in saving space for underground connections in busy cities.

Old and new communication carriers

Currently, a range of technologies provides our communications links, of which optical fibres are one component. Systems that have been used in the past were primarily designed to handle analogue audio signals. Copper cables were used in conjunction with regenerators or amplifiers, positioned regular distances apart, to strengthen the signal. Long-distance microwave links are also used for information transmission via a chain of towers in combination with microwave dishes and regenerators. Geostationary satellites have been used to provide international microwave links since the 1960s. These were originally designed for analogue multiplexed-voice or video signals.

To improve reliability and bandwidth, any new long-distance telecommunication systems only transmit digital information. Optical fibres are used as the backbone of most communication networks, with alternatives used for shorter links in order to minimise costs.

Other systems are still in place today. Although far more limited in terms of bandwidth than optical fibres, they are still useful. Smooth connections can be made between these other carriers and optical fibres.

Twisted-pair copper cables are suitable for local loop or simple telephone systems. A 600-pair cable can carry 600 two-way conversations. These do not have high bandwidth requirements, and because they operate over relatively short distances, attenuation stays at an acceptable level. The great advantages of twisted-pair cables are their low cost and simplicity of connection.

Coaxial cable offers a significant increase in bandwidth over twisted-pair cable, particularly through the use of multiplexing—sending analogue signals of differing wavelengths along the cable. Two cables of 50 tubes each are able to carry some 2700 two-way conversations. Coaxial cable is used over moderate distances with repeaters spaced every 45 km. It is lower in cost than optical-fibre cabling.

Microwave surface links are used over short distances, such as in wide-area computer networks (WANs). Approximately 2000 two-way conversations can be carried by a system with a repeater tower every 50 km.

Microwave satellite links are used for remote areas and for mobile communications. Their main disadvantage, apart from cost, is signal delay.

Telecommunications networks

The three main categories of telecommunications networks are local loops and local area networks (LANs), interoffice traffic networks and long-haul traffic networks. Balancing cost with service, optical fibres have already superseded copper in the interoffice and long-haul networks. Some LANs, such as the networks of university campuses and most schools, use optical fibres to connect host computers to associated terminals.

Physics file

SONET and SDH are sets of standards for the transmission of data over a fibre-optic network. SONET (Synchronous Optical Network) is used throughout America, whereas SDH (Synchronous Digital Hierarchy) is used in the rest of the world. The current standard maximum data rate, specified by SONET, is $2.488 \text{ Gbit s}^{-1}$.

Presently, while most Australian homes are not far from the fibre-optic superhighway, the connection to the home is made largely with copper cable. The biggest hurdle for linking homes directly with optical fibres is the high cost of the hardware needed to split the signal and convert it to a form suitable for home use. To reap the full benefits of optical-fibre technology and broadband capabilities, eventually local loops should also use fibre optics.

In the previous section you learned that optical receivers, placed at regular intervals, regenerate optical signals. Currently, these devices convert the light signals into electrical signals and then back into stronger light signals. The maximum bandwidth that can be achieved using optical fibres is limited by these electrical parts of the link. In the future, it is envisaged that these devices will be all optical; that is, the light signal will be enhanced without the need to convert it to an electrical signal. This will mean that repeaters will be more efficient and reliable, they will be placed further apart and eventually be lower in cost. As the cost decreases and more components are produced to provide links that are optical only, optical fibres will enter the local loop to provide connections to the home.

Choice of fibre

Multimode fibre, which suffers from modal dispersion, is generally used over short distances (up to about 2 km) with an LED as the light source. This fibre is also used over intermediate distances (up to 15 km) with laser diodes that generate light at a wavelength of 1310 nm for minimum attenuation. Laser diodes, only a few millimetres in length, are preferred as light sources to the larger gas lasers.

Single-mode fibre, which has a much better bandwidth but is more expensive, is used extensively for long-haul systems, with light at wavelengths of 1310 or 1550 nm produced by a laser diode.

The future of fibre-optic telecommunications

Engineers are researching techniques to increase the bandwidth of optical fibres even further. One key method that has been successful is *wavelength-division multiplexing* (WDM). Basically, many light signals at slightly different wavelengths are transmitted along one fibre. In this way the bandwidth is dramatically increased. At the moment, WDM can be achieved with 8, 16 or 32 wavelengths. It is imagined that it will eventually be possible to multiplex over 200 wavelengths onto a fibre.

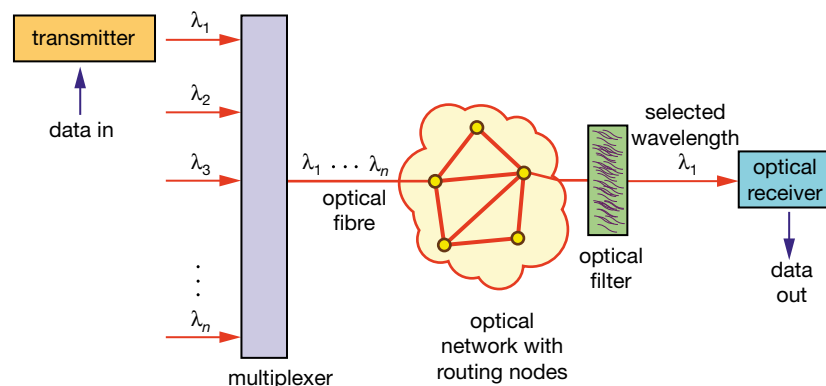


Figure 14.39 The set-up of a basic WDM system. Light of a number of wavelengths is combined by a multiplexer, the combined signal travels along an optical fibre and the required wavelengths are selectively tuned out at the other end by an optical filter. Each wavelength carries one signal.

As illustrated in Figure 14.39, light of a number of wavelengths is combined by a multiplexer and fed into the optical fibre. The signal is directed, or routed, through an optical network as required to reach the user. Selected wavelengths are then retrieved by means of a tunable optical filter at the receiver end. Before this system can become operational, a series of new optoelectrical devices must be invented, many of which are currently under development.

Optical fibre will eventually reach our homes. Unlike the older design, based on two-way conversations, the optical-fibre home connection will offer a broadband pathway for a range of options, from pay TV to home shopping. But for the time being, at least, the optical superhighway is still under construction!

Optical-fibre sensors

The rapid expansion of the telecommunications industry has led to the development of a vast array of state-of-the-art optical fibres and components. These advances have also produced many of the basic building blocks that are now used in a wide variety of fibre-optic sensing systems.

Fibre-optic sensors can be used in many applications where some of the following attributes are required:

- immunity from electrical interference
- chemical inertness
- compactness in size
- electrical safety.

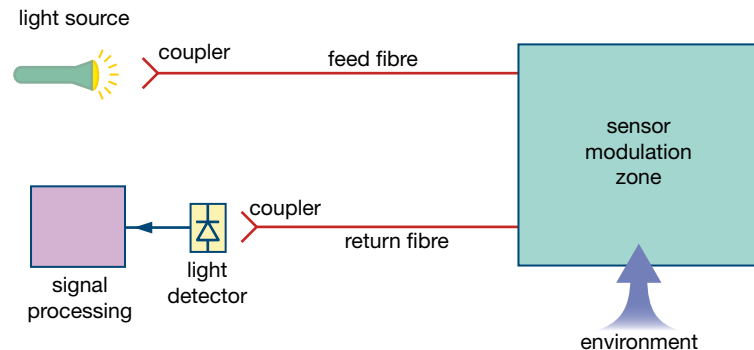


Figure 14.41 The functional blocks of a fibre-optic sensor system.

In this section, we shall explore a few different types of fibre-optic sensors used today. First we will discuss the basic operation of a fibre-optic sensor.

A fibre-optic sensor system

A fibre-sensor system can usually be described in terms of the basic functional blocks shown in Figure 14.41.

First of all, the sensor system requires a *light source*. This source needs to be bright, and most of the light energy needs to be emitted in one direction. The light source also needs to have a constant and stable intensity, frequency, phase or polarisation, or a particular combination of these, depending upon what property of the light is being modified by the sensor system. Usually an LED or laser diode (LD) is used as the light source.



Figure 14.40 Simple kits can be used to demonstrate the principals of WDM in the classroom.

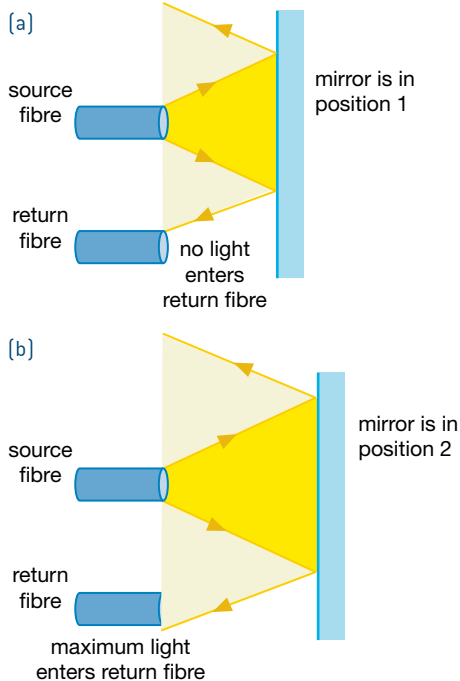


Figure 14.42 An extrinsic sensor used to monitor pressure. (a) The mirror is closer to the fibres, indicating high pressure. (b) The mirror is further from the fibres, indicating low pressure.

The light signal must then be *coupled* into the fibre core (with a diameter that can range from several hundred to less than 10 microns). This is usually achieved by some form of focusing lens system. Once the light is coupled into the core, it can propagate down the *feed fibre* to the *sensor modulation zone*. This is where the external physical parameter (temperature, pressure, position etc.) that is to be measured interacts with the light from the feed fibre. This parameter alters, or modulates, some property of the light (i.e. its intensity, phase, frequency or polarisation).

The modified light then travels to the detector via a *return fibre*. The light must again be *coupled* from the fibre to a suitable *detector*—usually a semiconductor photodiode. The value of the physical parameter under investigation is determined by monitoring the change in the selected property of the light. Some form of *signal processing* (such as demodulation, phase detection etc.) is often used to extract the relevant information from the return light signal.

There are two different types of sensor that can be used in the modulation zone: extrinsic and intrinsic sensors.

Extrinsic sensors

With an extrinsic sensor the modulation process is, as the name suggests, *external* to the fibre itself. Light is taken out of the feed fibre, modified in some manner, and then coupled back into the return fibre. Extrinsic sensors often make use of some form of light-intensity-attenuation process: the physical parameter that is being measured modulates the *amount of light* transmitted to the return fibre.

An example of an extrinsic sensor is a fibre-optic pressure sensor that uses a reflecting mirror to couple light between the feed and return fibres, as shown in Figure 14.42. The end faces of the two fibres are placed in close proximity to each other. Light from the feed fibre spreads out in a narrow cone from the fibre end and is reflected back towards the return fibre by a flat mirror attached to a flexible diaphragm that moves in response to external pressure variations. As the pressure changes, the mirror moves from position 1 to position 2, as shown in Figure 14.42. In the situation of maximum external pressure (Figure 14.42a), very little of the reflected light illuminates the core of the return fibre. As a consequence, the light level detected by the sensor system is at a minimum. In the situation of minimum external pressure (Figure 14.42b), the reflected light illuminates the entire core of the return fibre. In this case, the light level detected by the sensor system is at its maximum.

Intrinsic sensors

In an intrinsic sensor, the physical parameter being measured interacts directly with the light while it is inside the fibre. The advantage of this sort of system is that it does not require an interface coupling the light into the modulator and back into the return fibre. The disadvantage is that the physical parameter being measured may interact with the light not only in the modulation zone but also in the feed and return fibres. This can be a source of serious measurement error. The error can usually be minimised either by ensuring that most of the fibre is in the modulation zone, or by

altering the fibre slightly so that interaction occurs only in the modulation zone (as in the first example below).

An intrinsic fibre-optic sensor can be used to detect oil leaks in large oil tankers. The sensor consists of a fibre with a glass core surrounded by a plastic cladding. The glass core has an index of refraction of approximately 1.45 and the plastic cladding has a lower refractive index, ensuring that light propagates through the fibre core by total internal reflection. Part of the cladding is removed to expose the core—this becomes the modulation zone. When the exposed core is surrounded by water in the ballast tank (with a refractive index of around 1.33), light is totally internally reflected from the boundary between the glass core and the water, and the intensity of light transmitted through the fibre is unchanged. If there is any oil contamination in the water, oil droplets will attach themselves to the exposed glass core. The index of refraction of oil is approximately 1.5, and hence light will be transmitted (not reflected) from the core–oil boundary and be lost from the fibre (see Figure 14.43). The amount of attenuation of the light at the end of the return fibre is a direct measure of the oil concentration in the water. This type of detector is sensitive to oil contamination down to a level of a few parts per million.

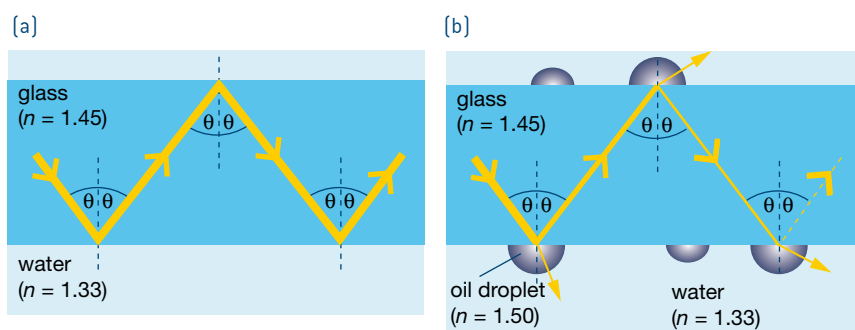


Figure 14.43 An optical fibre without cladding, used as an intrinsic sensor for the oil concentration in water. (a) When the fibre is surrounded by water, light is totally internally reflected along the fibre. (b) Contamination of the water by oil will result in some refraction of the light.

In section 14.3 you learned about microbending: due to small kinks in the fibre, some of the transmitted light is lost into the cladding. An intrinsic sensor can make use of this reduction in signal strength, or attenuation, to gauge the effect of some physical parameter that is made to deform the fibre itself. This technique can be seen in Figure 14.44.

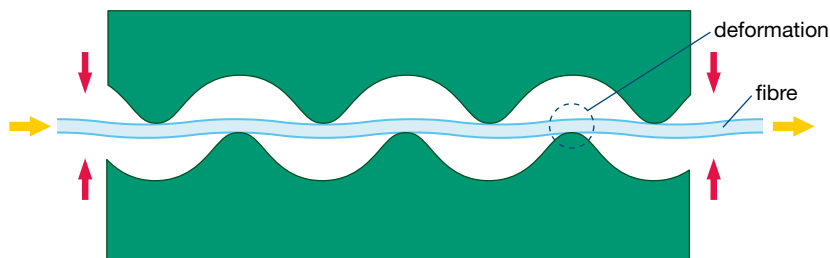


Figure 14.44 A sensor for measuring pressure or displacement. As the upper 'comb' structure is pushed down relative to the lower one, the optical fibre is deformed, resulting in some of the signal being lost due to microbending. The fluctuations in signal intensity reveal information about the pressure or displacement causing them.

All three sensors discussed in this section are examples of intensity-modulated fibre-optic sensors. Other properties of light that can be modulated include its frequency, its phase and its modal pattern.

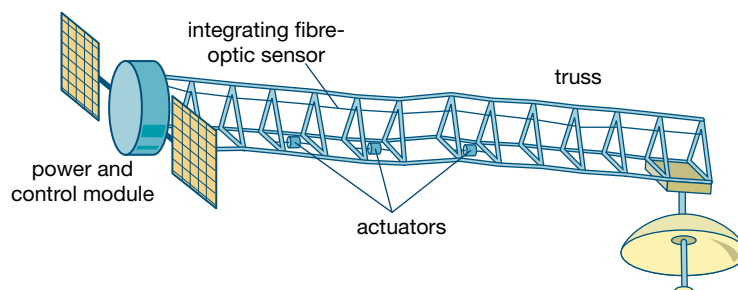


Figure 14.45 In satellite design, trusses are used to keep the power and control sections separate from the transmission sections. Because they are sensitive to heating and cooling, the trusses can be unstable. 'Smart' structures may one day solve this problem. Fibre-optic sensors positioned on the truss can detect temperature variations. They can then control actuators which actively dampen the effects of thermal changes.

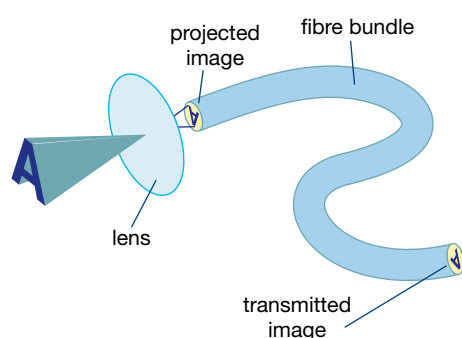


Figure 14.46 A lens and fibre bundle can be used together to produce a magnified transmitted image of an object.

Fibre-optic imaging bundles

As you know, light is channelled through an optical fibre almost without loss. It can follow a path with bends and turns, thanks to total internal reflection within the fibre. This principle can be applied to send an image from an area that would otherwise be inaccessible, in a way that is much more efficient than using a series of lenses and mirrors.

By binding a large number of individual optical fibres together in such a way that their arrangement is kept constant, a fibre-optic bundle is produced. This bundle can be used to transmit an image.

A fibre bundle is made up of a large number of individual optical fibres, arranged in an ordered array. Heating a bundle of individual optical fibres to a very high temperature fuses the cladding, so that the cores make up an ordered fibre bundle. The fused bundle is then carefully drawn out into a thin, flexible glass thread. The resulting miniature fibre bundle has a diameter of around one-quarter of a millimetre and contains approximately 10 000 glass cores, each with a diameter of a few microns and surrounded by a thin glass cladding. These fibres form a 10 000-pixel ordered array with good resolution.

Depending on the purpose, bundles may be manufactured with different end configurations: circular, square or rectangular. The most efficient way to pack the fibres for a circular end is in a hexagonal array.

It is possible to direct light of a particular wavelength along some fibres of a bundle and then have the light from interactions, such as fluorescence, return through different—or the same—fibres in the bundle. This makes fibre-optic imaging bundles a powerful tool for applications such as remote sensing.



Figure 14.47 Single-mode optical fibre terminated with an FC-type connector.

Microscopy in a tight spot

Medical researchers are developing special diagnostic dyes to detect the presence of individual cancerous cells. These dyes are preferentially taken up by the cancerous cells, which can then fluoresce (emit light of a unique colour) when exposed to UV illumination. Such a diagnostic aid can be used to detect very small numbers of cancerous cells well before they develop into a noticeable tumour. Some day it may help in the early detection and successful treatment of many potentially life-threatening cancers.

Fluorescing cells can easily be observed under a microscope, but this means that the cells have to be close to the microscope's objective lens. Hence, the cells have to be removed from the body as a biopsy and prepared on a microscope slide. Alternatively, if the cells are located near a surface inside the body, a high-magnification endoscope can sometimes be used to image the cells. But what if we want to observe fluorescence from individual cells deep within living tissue? Photonics researchers are currently working on a solution.

One end of a fused-fibre bundle is polished at a slight angle, so that it can be pushed into living tissue just like a hypodermic needle. The other end of the bundle can be viewed under a microscope. In this way, the probe is able to produce a magnified image of the piece of tissue in contact with the angled face of the bundle. The fused-fibre bundle does minimal damage as it penetrates the tissue, and it allows medical researchers to study the fluorescence of individual cells deep within living tissue.

These fibre bundles give good results when the end face is totally illuminated by a UV source, but the contrast of the fluorescence image is often poor. Recent studies have indicated that better contrast is achieved if the fibre bundle from is illuminated with a focused UV laser beam, which scans across

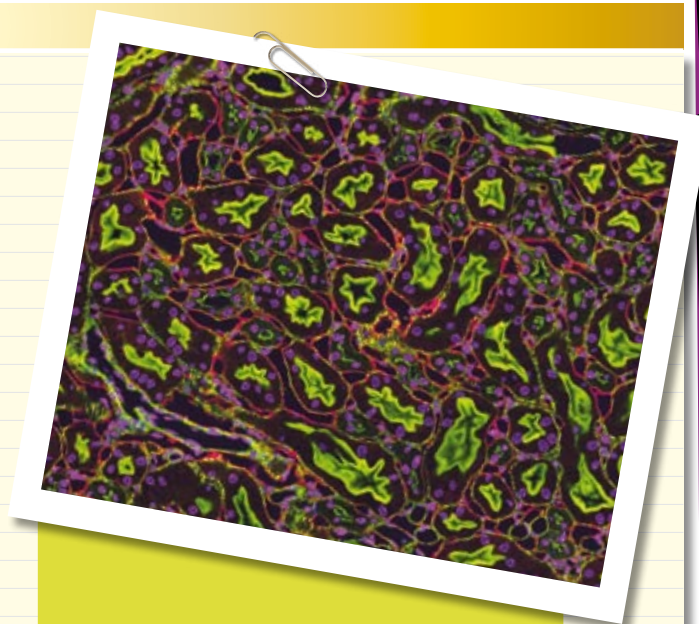


Figure 14.49 Kidney tubules, stained with a fluorescent dye, are imaged with a fused-fibre-bundle probe.

the bundle's surface. With this technique, only one pixel is illuminated at each particular instant and only the signal returning from this one pixel is detected. This eliminates much of the interference fluorescence coming from different cells. This is but one example of the wide applications of fibre-optic imaging bundles.

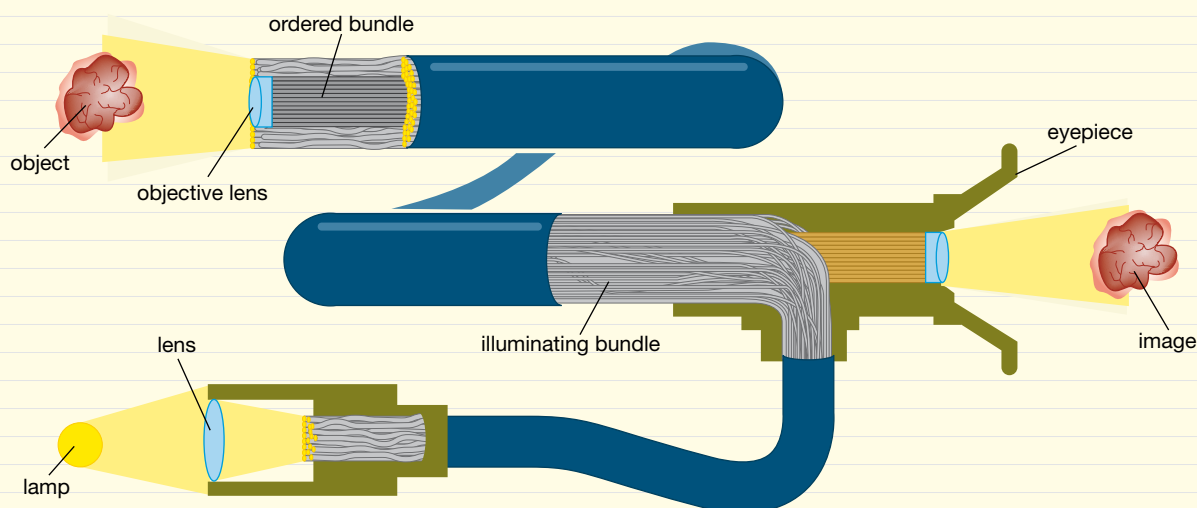


Figure 14.48 An endoscope, or fibroscope, is used to view inaccessible regions of the body. It typically consists of two non-ordered fibre-optic bundles that serve as light guides and one ordered bundle that produces the image.



14.4 summary

Applications of optical fibres

- Currently most intermediate and long-distance telecommunications links are made through optical-fibre networks. These work in combination with the existing twisted copper cables, coaxial cables, microwave surface links and microwave satellite links. The implementation of optical fibre to the home is likely in the future.
- For cost reasons, multimode fibres are used over short distances with LEDs and over intermediate distances with laser diodes as light source. Single-mode fibres are used over long distances with laser diode sources.
- The bandwidth of optical fibres may be increased by multiplexing: a number of signals at different wavelengths are sent through the one fibre in a process called wavelength-division multiplexing (WDM).
- Extrinsic sensors often use some form of light-intensity-attenuation process. The amount of light transmitted to the return fibre is modulated by the physical parameter that is being measured.
- In the case of an intrinsic sensor, the physical parameter to be measured interacts directly with the light while it is inside the fibre.
- 'Smart sensors' is a rapidly developing field, in which structures are embedded with optical fibres that can sense changes in their environment and then react in some manner.
- A fibre-optic bundle is produced by binding a large number of individual optical fibres together in such a way that their arrangement is kept constant. This bundle can be used to transmit an image.



14.4 questions

Applications of optical fibres

- 1 Given that 94.5 kg of copper wire is required to perform the same task as 3.6 kg of optical fibre, calculate the percentage weight reduction involved in converting a copper network to an optical-fibre network.
- 2 Explain why optical fibres are extensively used in interoffice and intercity traffic networks, but are not generally found in LANs or in the local loop.
- 3 Twisted copper cable, coaxial cable and microwave surface links are all used effectively in communication systems. What is the key advantage of an optical network?

The following information relates to questions 4 and 5. A multimode fibre being used for a short-distance telecommunication link has a Ge-doped core with a refractive index of 1.460. The cladding surrounding the core has a refractive index of 1.440.

- 4 Calculate the numerical aperture for the system, given that the fibre operates in an air-filled room.
- 5 Suggest a reason why a multimode fibre is used for this application rather than a single-mode fibre.
- 6 Refer back to the illustration of a moving-reflector sensor in Figure 14.42. A movement of the mirror of

less than 200 μm can make the light output change from zero to the maximum level. Explain how optical fibre is used in this equipment to act as a motion or pressure sensor.

- 7 Explain the difference between an extrinsic and an intrinsic fibre sensor. Give an example of each.

The following information relates to questions 8 and 9. Recall the example of the sensor used in large oil tankers to monitor whether oil is leaking into the ballast-water tanks. A particular sensor consists of a fibre with a glass core of refractive index 1.45, surrounded by a plastic cladding of slightly lower refractive index. Part of the cladding is removed to allow the fibre core to have contact with the external water, which has a refractive index of 1.33.

- 8 Calculate the critical angle of the core-water interface.
- 9 Given that the oil has a refractive index of approximately 1.50, explain how the sensor is able to detect any oil leaks that may occur.
- 10 Looking at the diagram of the endoscope shown in Figure 14.48, explain the difference in operation of the ordered and non-ordered light guides.



chapter review

Use the following values where needed:

$$h = 6.63 \times 10^{-34} \text{ J s} = 4.14 \times 10^{-15} \text{ eV s}$$

$$c = 3.00 \times 10^8 \text{ m s}^{-1}$$

$$hc = 1240 \text{ eV nm}$$

Multiple-choice questions

1 Use the following key to answer this question.

- A** incandescent light bulb **B** sodium vapour lamp
C light-emitting diode **D** laser
E none of the above

The light of which is:

- a** generated by the electrons in individual atoms as they drop one or more energy levels?
b generated by random thermal motion of atoms in the material?
c generated as electrons fall from the conduction band to the valence band in a semiconductor?
d coherent?

2 Use the following key to answer this question.

- A** infrared **B** visible
C ultraviolet **D** X-ray

All objects radiate electromagnetic radiation, but the nature and amount of this radiation changes with their temperature. Refer to Figure 14.2, page 510.

- a** Which type of radiation is dominant in the spectrum from an object at room temperature?
b The temperature at the Sun's surface is approximately 6000 K. What is the dominant radiation in sunlight?
c The hottest stars have temperatures around 25 000 K. In what range will the majority of their radiation lie?

3 Infrared LEDs are often used in remote controls for TV sets. What can you say about the gap between the valence and conduction bands in the semiconductor material of these LEDs?

- A** The other wavelengths of light within the room must all lie outside the infrared energy range.
B The band gap must be more than the energy of a photon of light within the infrared range
C The band gap must be less than the energy of a photon of light within the infrared range
D Nothing. I need more information!

4 Red LEDs have been around for a long time, but white LEDs have been developed only recently. Why has it been difficult to produce them and how are they made?

- A** LEDs produce light only of one specific colour. White LEDs package other colours to produce near-white light.
B White light is produced by high-energy photons. White LEDs require high-energy batteries only recently available.

- C** White light is produced by absorbing all other colours. Suitable filters need a very small band gap
D All of the above.

5 What energy gap should a semiconductor material have if it is to emit visible light?

- A** Less than 1.5 eV
B Between 1.5 and 2 eV
C Between 2 and 2.5 eV
D More than 2.5 eV

6 What is required if an incandescent light bulb is to produce coherent light?

- A** The filament has to be very tiny.
B The filament has to be made of a doped-silicon semiconductor.
C The bulb has to be very highly evacuated.
D It is impossible for a light bulb to produce coherent light.

7 Which one or more of the following are true of laser-light photons?

- A** They all have the same wavelength.
B They are all in phase.
C They all have the same energy.
D They all travel at the same speed.
E All of the above

Extended-answer questions

8 A ruby laser produces light with a wavelength of 690 nm. It is used to produce short powerful pulses of light, lasting only 2.0×10^{-11} s. The power of each pulse is 5.0 GW.

- a** How many photons are released in each pulse?
b How long (in m) are the pulses?
c How many photons are there in each millimetre of pulse?

9 An atom is stimulated by a photon and releases a photon of equal energy to the photon that stimulated it, and yet the original photon does not lose energy. Why is the law of conservation of energy not violated?

10 In a laser, the action takes place in a 'cavity' which has mirrors at either end. What is the purpose of the mirrors and how does the light escape from the cavity?

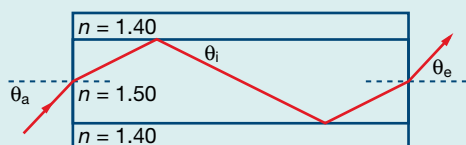
11 Laser light is used to read DVDs and to 'weld' blood vessels in the retina of the eye, because it can be focused to a point more precisely than light from a normal light source. Why is this?

12 Light signals, whether travelling in optical fibres or simply in air, can carry a much greater amount of information than signals in copper wires. What is the fundamental reason for this?

13 There are various reasons why fibre-optic cables are never used in air without any cladding. One reason is that they are

delicate and would easily be broken. There are more fundamental reasons, however. What are they?

- 14 The cladding of an optical-fibre cable has a slightly lower refractive index than the core. Why is this, and why should the cladding not have a much lower refractive index?
- 15 Why is optical-fibre cabling not used to carry the signals right into our homes?
- 16 The diagram shows a simple optic light pipe. The core has a refractive index of 1.50 and the cladding an index of 1.40.



- a What is the maximum angle, θ_p , for which the light will be totally reflected?
 - b What is the acceptance angle, θ_a , the maximum angle at which light entering the pipe will be transmitted down its length?
 - c At what maximum angle, θ_e , does the light emerge at the far end?
 - d Proportionately, how much further will light travel if it zig-zags down the length of the pipe at this angle than if it travels straight down the centre?
 - e What is the numerical aperture of this fibre?
- 17 Step-index multimode fibres and graded-index multimode fibres usually have about the same diameter, and yet the latter has much less modal dispersion. Describe both types of fibre. Why does graded-index fibre have less modal dispersion?
 - 18 Despite the fact that a cable of larger diameter can accept more light, long-distance optical-fibre cables are generally narrower than those used in short-distance applications. Why is this? What are the advantages of each type?
 - 19 An optical fibre has a core with a refractive index of 1.52 and a cladding with a refractive index of 1.49.
 - a What is the numerical aperture for this fibre?
 - b Why do you think usually the numerical aperture, rather than the acceptance angle, of an optical fibre is quoted?
 - c Calculate the acceptance angle for this fibre when:
 - i it is in air
 - ii it is immersed in water ($n = 1.33$).
 - 20 Ultraviolet light has a higher frequency than infrared light, so it would be capable of transmitting information at a higher frequency. Why is it, however, that almost all optical-fibre technology uses infrared rather than ultraviolet light?
 - 21 The modal dispersion in a 1 km length of multimode optical fibre is said to be around 100 ns, while in a single-mode fibre it is only 1 ns.
 - a Using these figures, what would be the approximate maximum frequency at which data could be transmitted along a 20 km length of single-mode fibres?
 - b If it were necessary to transmit data at 1 GHz, what would be the maximum length of each of these fibres that could be used?
 - 22 The refractive index of crown glass varies from 1.532 for violet light to 1.513 for red light. Explain (with appropriate calculations) why if white light pulses were used to carry data down a crown-glass fibre, the rate at which the data would be transmitted would be limited to around 10 MHz for a kilometre of cable. What is the name for this type of limit on the transmission of data?
 - 23 Which one or more of the following can cause attenuation of the signal in a fibre-optic cable?
 - A Too great a difference between the refractive indices of the core and the cladding.
 - B Scattering of the light by the atoms of the glass.
 - C Absorption of the light by impurities in the glass.
 - D Slight variations in the refractive index of the core.
 - E Too many sharp bends in the cable.
 - 24 The attenuation of the signal in a fibre is measured in decibels. An attenuation of -3 dB means a drop of:
 - A 3 W of power
 - B half the power
 - C 30% of the power
 - D 3% of the power.
 - 25 Infrared light used in optical fibres often has a wavelength of 1300 or 1500 nm, but not 1400 nm. Why are these particular wavelengths chosen?
 - 26 Both LEDs and lasers can be used as a source of IR light to carry signals in optical fibres. While LEDs are used for relatively short stretches and lower-speed systems, lasers need to be used for long-distance high-speed connections. Explain the reasons for this.
 - 27 Photonics engineers have developed 'optical repeaters', devices which amplify the optical strength of a signal without converting it to an electrical signal. On the other hand, 'regenerators' are still required at regular intervals along a long system. What are the differences between optical repeaters and regenerators?
 - 28 Optical fibres can be used as sensors in a variety of ways. In a fibre-optic sensor device the light is:
 - A always modified within the fibre and returned to a detector for analysis
 - B always modified outside the fibre and then returned into the fibre
 - C modified either inside or outside the fibre
 - D modified both inside and outside the fibre.
 - 29 Describe some of the ways in which light can be modified in a system that uses optical fibres to detect or measure a quantity.
 - 30 An endoscope is a medical device that contains one or two unordered bundle[s] of optical fibres surrounding an ordered bundle. Describe how this device can be used to observe inaccessible regions inside the body.



Sound



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 form 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.

An investigation of the recording and reproduction of sound starts with a good understanding of sound, but we also need to examine the means by which we humans can mimic the natural world. Our microphones and loudspeakers do no more than the human ear and voice, relying on electrical and magnetic interactions rather than biological structures to amplify or modify the original sound.

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.



by the end of this chapter

you will have covered material from the study of sound, including:

- the longitudinal wave nature of sound (pitch, frequency and period, amplitude, wavelength, speed)
- intensity, sound levels and loudness
- the diffraction of sound through gaps and around obstacles
- the superposition of waves from more than one source
- standing waves and resonance
- frequency response curves for human ears, speakers and microphones
- the detection and recording of sound by the ear and various types of microphones
- the production of sound by musical instruments and loudspeakers.



outcome

On completion of this chapter, you should be able to apply a wave model of sound and ideas about electromagnetism to describe, analyse and evaluate the recording and reproduction of sound.



CHAPTER 15

15.1 The nature of sound

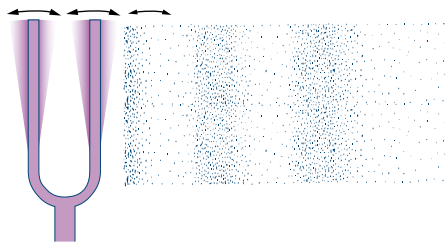


Figure 15.1 The rapid vibration of a tuning fork makes the air molecules around it vibrate in much the same way as the human larynx or a loudspeaker—creating the compressions and rarefactions of a pressure wave that we perceive as sound.

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.

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.



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 15.2 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.

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?



Figure 15.3 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.

We'll see later how the loudspeaker works, but consider for the moment what it's doing. 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 15.4 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.

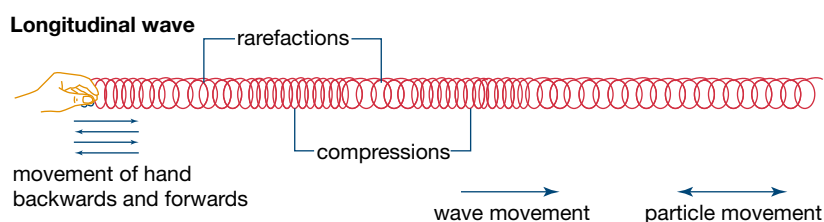


Figure 15.4 The transfer of energy in a spring by successive compressions and rarefactions, i.e. a longitudinal wave.

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*.



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.

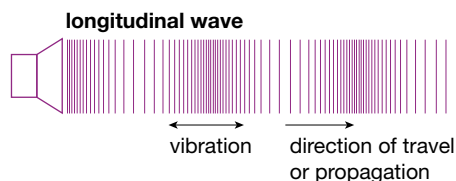


Figure 15.5 In a longitudinal wave, the vibrations of the individual particles are in the same direction as the direction of wave propagation.

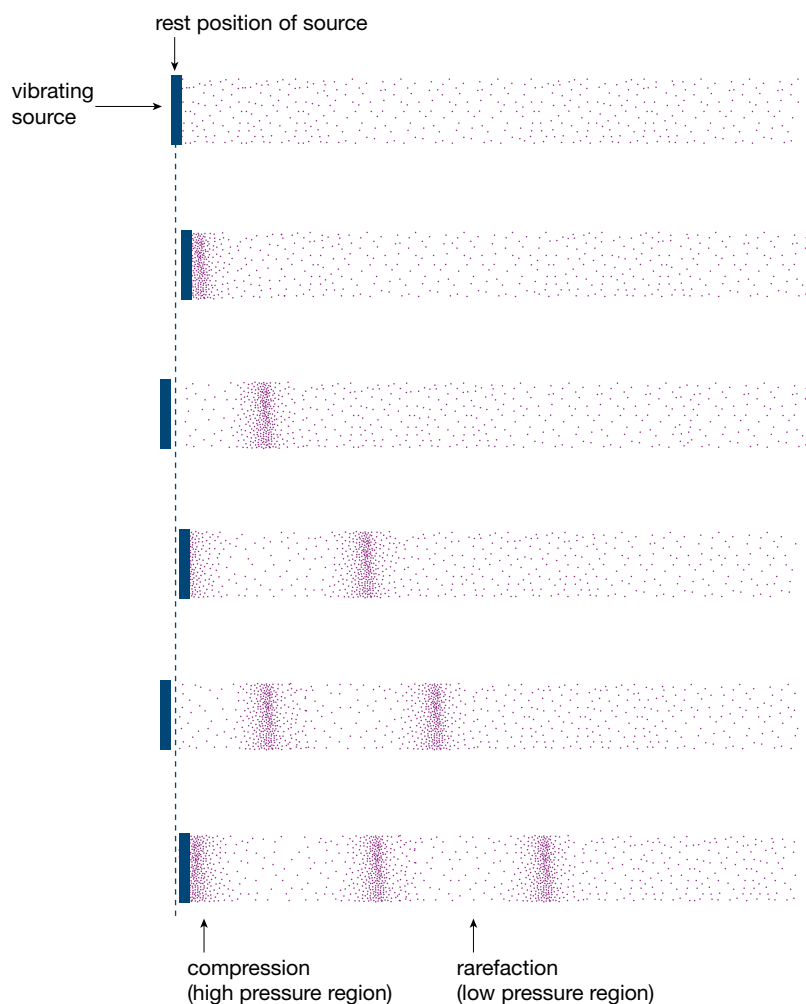


Figure 15.6 In a medium in which there is no sound, the particles are evenly spread. As the source moves backwards and forwards, the continual vibration produces a series of compressions and rarefactions moving continuously away from the source. In this way, sound energy is propagated through a medium. 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 15.7). 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.

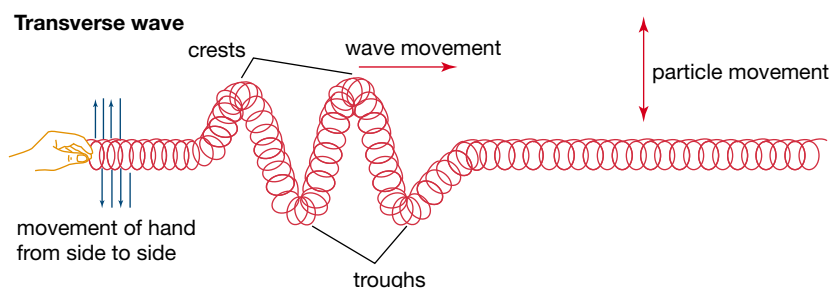


Figure 15.7 The transfer of energy in a spring by transverse movement.

Transverse wave

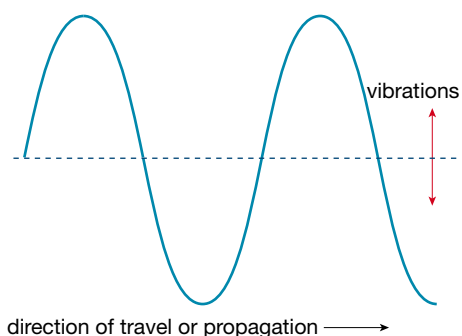


Figure 15.8 In a transverse wave, vibrations are at a right angle to the direction of travel of the wave.

Representing sound as waves

Picturing sound as a series of compressions and rarefactions in the medium in which it moves makes sense, but it's 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 15.9 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.

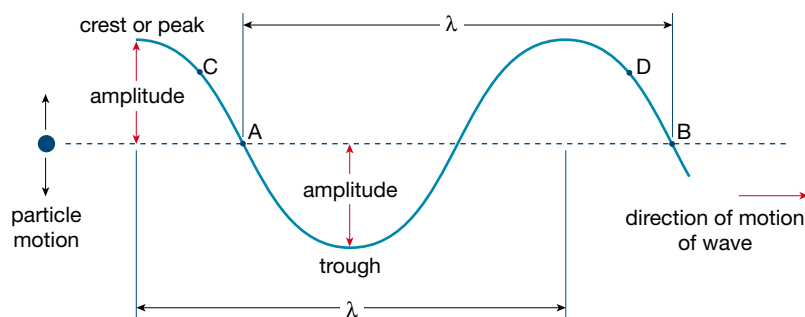


Figure 15.9 A simple pictorial representation of a transverse wave.

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

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.

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). The *period*, T , is the time taken for one cycle to be completed in seconds (s), hence:

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

For example, if 10 crests pass a given point in 1 second, then the frequency is 10 cycles per second, or 10 Hz. The period, or time for each complete wave, will be $1/10$ seconds, or 0.1 s.

Since these definitions apply to all waves, let's examine how they apply to longitudinal sound waves. In Figure 15.10a, 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 15.9. 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 15.10b. 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.

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 waiting surfers sitting on their boards.

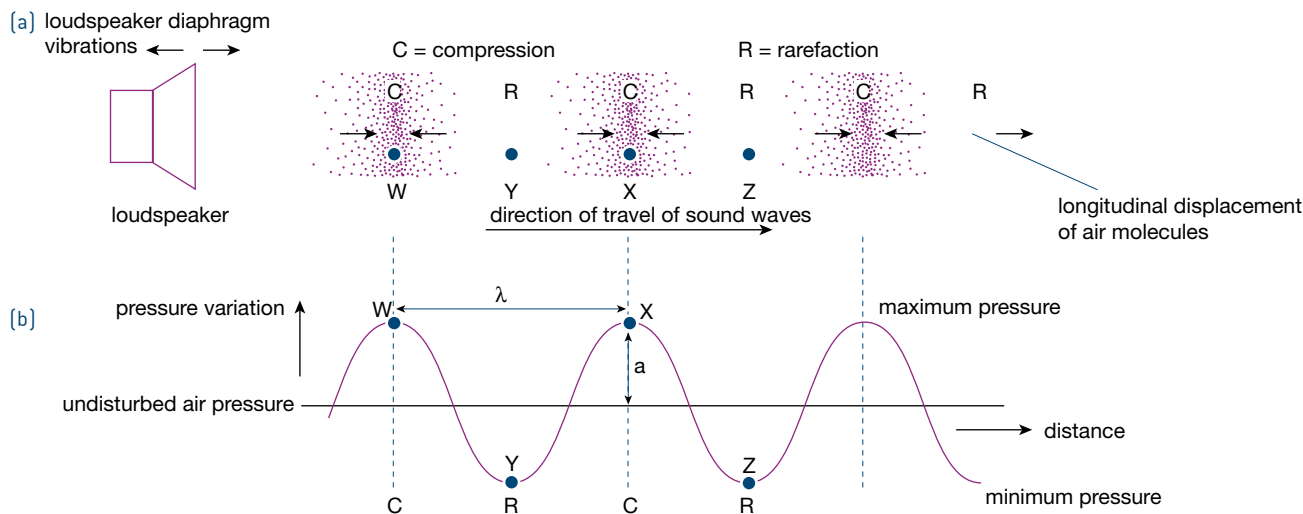
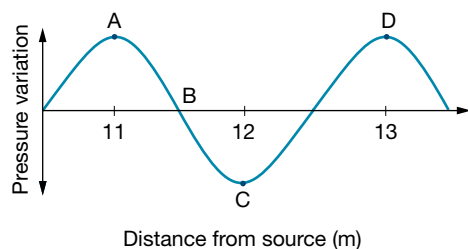


Figure 15.10 A longitudinal wave as a simple pictorial representation.



Worked example 15.1A

Sound is transmitted through air as a longitudinal wave. The graph on the left represents the pressure variation of a sound wave of a single frequency at a particular instant.

- What is the wavelength of the sound wave?
- Which of the points A–D represent compressions?

Solution

- 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:

$$\lambda = 13 - 11 = 2 \text{ m}$$
- Points of compression are represented as points of maximum positive pressure; hence, A and D are currently at points of compression.

Seeing sounds

A sound wave is not easy to visualise, but you can get an idea of the characteristics of a particular sound wave by digitising the sound wave and viewing it by using suitable computer software or by using a cathode ray oscilloscope (CRO) if more recent equipment is not available. Figure 15.11 illustrates the pattern produced when a loudspeaker connected to a signal generator is placed near a microphone connected to a CRO or computer. The steady note from a signal generator produces a uniform trace that looks exactly like a transverse wave, but we know that the sound producing it is a longitudinal wave. Crests and troughs on the screen correspond to compressions and rarefactions in the sound wave (so the graph actually represents pressure variation with time).

Figure 15.11 also illustrates the effect of changing the loudness and frequency on the trace. Increased loudness, which increases the energy of the sound wave, results in an increased amplitude for the trace on screen. As the frequency of the signal is increased, the number of full cycles, seen as wavelengths on the screen, will increase.

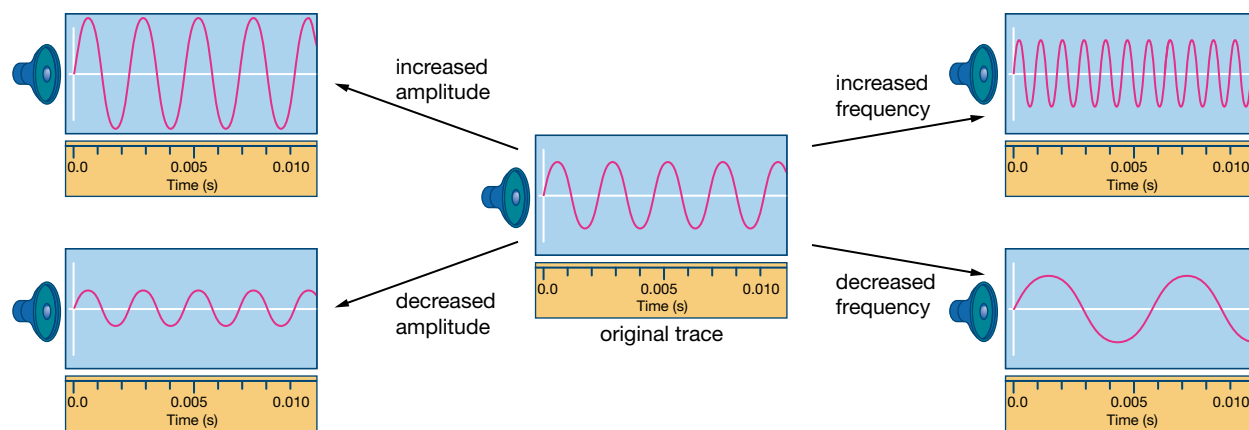


Figure 15.11 By using digitising software or a CRO, a longitudinal sound wave can be viewed as a transverse graph. The trace shows how the air pressure in the wave changes with time.



15.1 summary

The nature of sound

- Sounds are produced by the vibration of objects and require a medium in which to travel. Sound cannot be transmitted in a vacuum.
- All forms of wave motion transfer energy without a net transfer of matter. Once a sound wave passes, the particles of the medium return to their original position.
- Sound is a longitudinal wave. Vibrations occur in the same direction as the direction of propagation of the wave.
- 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).
- 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.



- The wavelength, λ (m), of a wave is the minimum distance the wave travels for one complete cycle, also defined as the distance between successive points 'in phase'.
- The frequency, f (Hz), of a sound wave is the number of waves that pass a point in 1 second.
- The period, T (s), of a sound wave is the time for one complete cycle to pass a point: $f = \frac{1}{T}$.

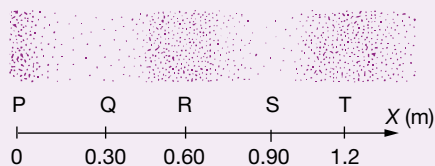
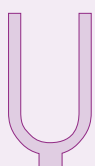
- The amplitude, A (m), of a sound wave is the maximum displacement of a particle from its undisturbed position. A larger amplitude requires more energy, so a louder sound means a larger amplitude.
- A sound wave may be represented by a graph of pressure variation vs. distance from the sound source.



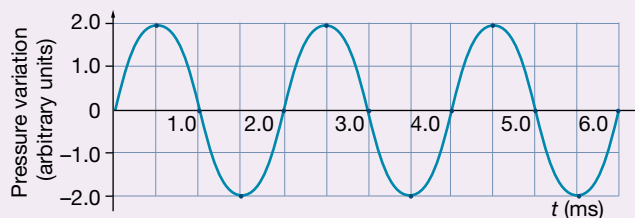
15.1 questions

The nature of sound

- Which one of the following correctly describes the reason we can hear sound from a vibrating tuning fork?
 - The air molecules adjacent to the prongs are propelled towards our ears and eventually strike the eardrum.
 - The kinetic energy of the vibrating prongs is transmitted instantaneously to the eardrum in the same way that electromagnetic energy (or light) is transmitted.
 - The prongs produce vibrations in adjacent air molecules, which results in the transfer of energy to other adjacent molecules and eventually to the eardrum.
- Which one or more of the following statements is true?
 - Sound can travel in a vacuum.
 - During the transmission of sound, air molecules are permanently displaced from their original positions.
 - Sound exhibits wave-like behaviour.
 - The transfer of sound energy is independent of the medium.
 - The transmission of sound involves only the transfer of energy and not matter.
- Which of the following describes the direction of the vibrations of the air molecules in the following wave?
 - \leftrightarrow
 - \rightarrow
 - \downarrow
 - \uparrow
 - \leftarrow



- For the wave in Question 3, which of the following alternatives correctly describes the direction of energy transfer of the sound between the tuning fork and point X?
 - \leftrightarrow
 - \rightarrow
 - \downarrow
 - \uparrow
 - \leftarrow
- Describe the difference between transverse and longitudinal waves on the basis of the particle motion in the wave.
- Calculate λ for the wave in Question 3.
- A microphone is connected to a digitiser and placed at point X in the diagram in Question 3. The pressure variation vs. time trace displayed on the connected computer screen is as follows.



- Describe the type of energy transformation occurring at the microphone.
 - At what time, or times, did maximum pressure variation occur?
- For the wave in Question 3, describe the pressure variation at the following points:
 - P, R, T
 - Q, S
 - a point 0.15 m to the right of P
 - a point 0.75 m to the right of P.
 - From the graph shown in Question 7, when did:
 - a compression occur?
 - a rarefaction occur?

15.2 The wave equation

When a sound source is close by, it is easy to assume that you hear the sound as soon as it is produced. But other common experiences suggest that although sound travels very rapidly, its speed is considerably less than the speed of light. For example, you usually see lightning before you hear the thunder, and at an athletics meeting you can see the smoke of a starter's gun before you hear its sound.

Sound travels at different speeds in different media. How it behaves depends on the characteristics of the sound wave and the nature of the medium.

The speed of sound

The first known measurement of the speed of sound was made in 1640 by a French mathematician, Marin Mersenne. He simply banged two iron bars together to create a loud sound, and then measured the time it took for the sound to travel to a distant object and be reflected back to him. Knowing the total distance, he could calculate its speed from the relationship:

$$\text{speed} = \frac{\text{total distance travelled}}{\text{time}} = \frac{d}{t}$$

In air at room temperature, sound travels at about 340 m s^{-1} . The speed of sound is usually greater in liquids than gases, and greater still in solids. Table 15.1 summarises the speed of sound in various different materials at 20°C (except where shown).

It is important to understand that the speed of sound is the speed with which the vibration is passed from particle to particle. This is quite different from the speed of individual particles within the transmitting medium. For example, in air the molecules travel at 500 m s^{-1} , while the sound being transmitted travels at 340 m s^{-1} . The speed of the wave depends upon the forces within the medium pulling particles back to the mean position. In solids, this force is usually considerably greater than in liquids or gases, and so the wave travels at a greater speed.

Table 15.1 The speed of sound in different materials

Material	Speed of sound (m s^{-1})
Carbon dioxide	266
Air (0°C)	331
Air (20°C)	344
Lead	1300
Water	1410–1500
Mercury	1450
Rock	1500–3500
Brass	3500
Copper	3970



Figure 15.12 Knowledge of the speed of sound in air provides a rough means for calculating the distance from an observer to the centre of a thunderstorm. Each 3 s time interval between a lightning flash and the corresponding thunderclap represents approximately 1 km to the centre of the storm.



PRACTICAL ACTIVITY 47

Speed of sound by clap and echo

Factors affecting the speed of sound

The speed of a wave in a particular medium is related to the characteristics of the wave and the medium: different wave types will have different characteristic speeds. Sound travels at approximately 340 m s^{-1} in air, but the speed of sound can be quite different in other media and sound cannot travel at all in a vacuum. The size, mass and spacing of the particles within the medium all influence the speed at which the sound wave's energy is transmitted. In a liquid, where the molecules are closer together than in air, molecules need less movement to transfer the energy of vibration, and the speed of sound is greater. For similar reasons the speed of sound in solids is, generally, even greater.

A change in temperature of a gaseous medium will have a marked effect on the speed at which a sound wave will travel. In Table 15.1, different speeds are quoted for 0°C and 20°C . The speed of sound in air increases by approximately 0.6 m s^{-1} for every 1°C increase in temperature for a given volume of air. A higher temperature means that the air particles are moving faster, so they propagate the pressure changes more rapidly.

The speed of sound in air is not significantly affected by the frequency of the sound wave. This makes sense when you consider the sounds produced by various instruments in a band: the high frequencies of a flute or mouth organ reach your ears at the same time as the lower frequencies of the bass guitar.

The wave equation

Although the speed of sound can vary, there is a relationship between the speed of the wave and other significant wave characteristics. Since speed is given by:

$$v = \frac{\text{distance travelled}}{\text{time taken}} = \frac{d}{t}$$

The speed of any wave travelling a distance of one wavelength (λ) in one period (T), will be:

$$v = \frac{\lambda}{T}$$

where v = wave speed (m s^{-1})

λ = wavelength (m)

T = period (s).

And since $f = \frac{1}{T}$:

$$v = \frac{1}{T} \times \lambda = f \times \lambda$$

and the relationship becomes:

$$v = f\lambda$$

where v = speed (m s^{-1})

f = frequency (Hz)

λ = wavelength (m).

This is known as the *wave equation*. In a given medium at constant temperature, the velocity will be constant, so the wavelength is inversely proportional to the frequency (i.e. $\lambda \propto \frac{1}{f}$).

The wave equation also implies that for a sound of constant frequency, the wavelength will be affected by the velocity of the sound (i.e. $\lambda \propto v$). This explains why you must warm up a wind instrument before playing with a band or orchestra: sound travels at different speeds in warm and cold air, so a cold clarinet, for example, will not play exactly the same notes as a warm one.

Worked example 15.2A

Look at the sound wave shown for Worked example 15.1A. Assume that the speed of the sound wave is 340 m s^{-1} .

- a What is the frequency of the sound?
- b What is the period?

Solution

- a The wavelength [from point A to D, as calculated in section 15.1] = 2 m.
Rearrange the wave equation so that it reads $f = \frac{v}{\lambda}$. We know that $v = 340 \text{ m s}^{-1}$,
so $f = \frac{340}{2} = 170 \text{ Hz}$.
- b $f = \frac{1}{T}$, so that $T = \frac{1}{f}$
so $T = \text{or } 5.9 \times 10^{-3} \text{ s} = 5.9 \text{ ms}$

Worked example 15.2B

Sonar tests off the south coast of Tasmania found that the speed of sound in air was 340 m s^{-1} , while the speed of sound in cold sea water was $1.50 \times 10^3 \text{ m s}^{-1}$.

- a What is the wavelength of a 300 Hz sound made under the water?
- b What is the ratio of the wavelength of the sound wave in air to that of the same sound made under water?
- c The sound is repeated in air at twice the volume. What change does this make to the ratio of the wavelengths?

Solution

- a Rearrange the wave equation so that it reads $\lambda = \frac{v}{f}$. We know that
 $v = 1.5 \times 10^3 \text{ m s}^{-1}$ and $f = 300 \text{ Hz}$, so $\lambda = \frac{1.5 \times 10^3}{300} = 5.0 \text{ m}$
- b The ratio of wavelength in air will be the same as that of the ratio of the respective speeds since λ is directly proportional to v , so ratio = $\frac{v_1}{v_2} = \frac{340}{1500} = 0.23$.
- c The volume of a sound affects the energy of a sound wave, and hence its amplitude. It does not affect the wavelength, so the ratio will not change.



15.2 summary

The wave equation

- In air, sound travels at approximately 340 m s^{-1} , or about 1 km every 3 s.
- The speed with which a wave travels through a particular medium is related to the characteristics of the wave and to the size, mass and spacing of the particles in the medium.
- The speed of sound in air is influenced by the air's temperature. It increases by approximately 0.6 m s^{-1} for every 1°C increase in temperature.
- The speed of a sound wave in air is independent of the frequency.
- The wave equation links velocity, frequency and wavelength by the equation $v = f\lambda$.
- In a constant medium at constant temperature, the speed of sound is constant and $\lambda \propto \frac{1}{f}$.
- For a note of constant frequency, $\lambda \propto v$.



15.2 questions

The wave equation

- Which one of the following does the phrase 'speed of sound in air' refer to?
 - The speed at which individual air molecules travel from the source of the sound
 - The speed at which individual molecules vibrate while transmitting the sound
 - The speed at which the sound energy is transferred from the source
- Explain why you would expect sound to travel faster in water than through air.
- Which of the following has a significant effect on the speed of sound in air?
 - The frequency of the sound wave
 - The amplitude of the sound wave
 - The temperature of the air
 - The air pressure
- The energy carried by a sound wave is primarily a function of which of the following characteristics?
 - velocity
 - wavelength
 - speed
 - amplitude
- Which one of the following properties of sound is independent of the others?
 - period
 - frequency
 - amplitude
 - wavelength
 - speed
- What is the speed of the sound wave produced by the vibrating tuning fork shown in questions 3 and 7 at the end of section 15.1?
- At the start of an athletics sprint final, the starter is standing 10 m from the favourite in the inside lane and 20 m from the second favourite in the outside lane. If the favourite starts on seeing the flash on the starter's gun while the second favourite starts on hearing the sound of the gun, what advantage in time will the favourite gain over the second favourite? (Assume the speed of sound in air is 340 m s^{-1} .)
- An artillery shell fired at a target 3.4 km away was heard to strike the target 20 s after leaving the gun. Calculate the average speed of the shell, given that the speed of sound in air is 340 m s^{-1} .

The following information applies to questions 9 and 10. A scientific research ship used sonar to determine the ocean depth. A 1.0 kHz signal was bounced vertically off the sea bed and was received 2.0 s after transmission. The speed of sound in the water was 1.5 km s^{-1} .

- What was the wavelength of the signal as it travelled through the water?
 - Calculate the ocean depth at this location.
 - If this signal was also directed into the air, calculate the ratio of the frequency in air to that in water.
- The scientists decide to check the depth by using a 2.0 kHz signal of the same amplitude.
 - What is the wavelength of this new signal as it passes through the water?
 - Will the time interval between the transmission and reception of this new signal be the same as before? Justify your answer.

15.3 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, as we will see later, and comes about as a result of a wave phenomenon known as *diffraction*.

How much a particular wave spreads 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.



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.

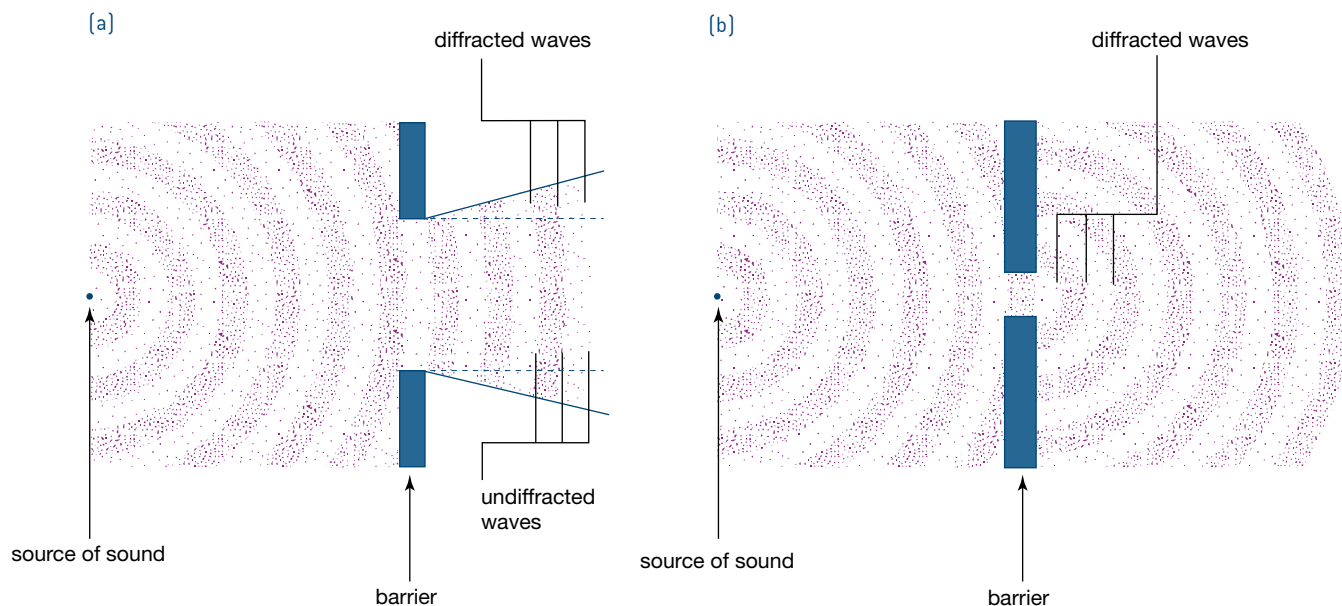


Figure 15.13 The amount of diffraction depends on the size of the gap, or aperture, and the wavelength. When the aperture is much larger than the wavelength, only limited diffraction will occur (a). Wavelengths larger than the aperture width will result in significant diffraction (b).

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

More on types of speakers and their roles in speaker systems can be found in section 15.7.

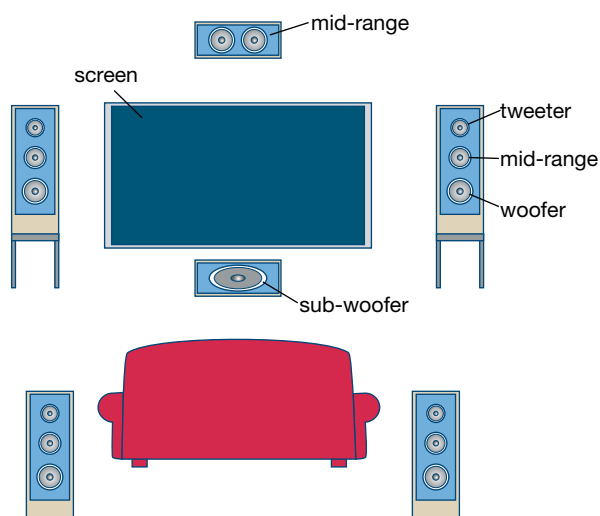


Figure 15.14 Most modern speaker systems use moving coil loudspeakers designed to reproduce particular ranges of frequencies: large base 'woofers', or even a single sub-woofer, for the lower non-directional frequencies (20–500 Hz), multiple mid-range units (500 Hz–4 kHz) and many directional 'tweeters' 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.

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.

Worked example 15.3A

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.

- Assuming that the speed of the sound is 340 m s^{-1} , calculate the wavelength of a 9000 Hz sound.
- Explain why this high-frequency sound leaves a sound shadow on one side of a person's head.

Solution

- a** $v = 340 \text{ m s}^{-1}$ and $f = 9000 \text{ Hz}$.

$$\text{Using the wave equation } v = f\lambda, \text{ or } \lambda = \frac{v}{f} = \frac{340}{9000} = 0.0378 \text{ m}$$

i.e. the wavelength is 0.0378 m or 3.8 cm.

- b** A person's head is about 20 cm in diameter. The wavelength of a 9000 Hz sound (3.8 cm) is significantly smaller than this, and since diffraction is only significant when λ/w is approximately 1 or more, diffraction will be minimal and the sound will not bend significantly around the head.

Worked example 15.3B

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.



15.3 summary

Diffraction of sound

- 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.

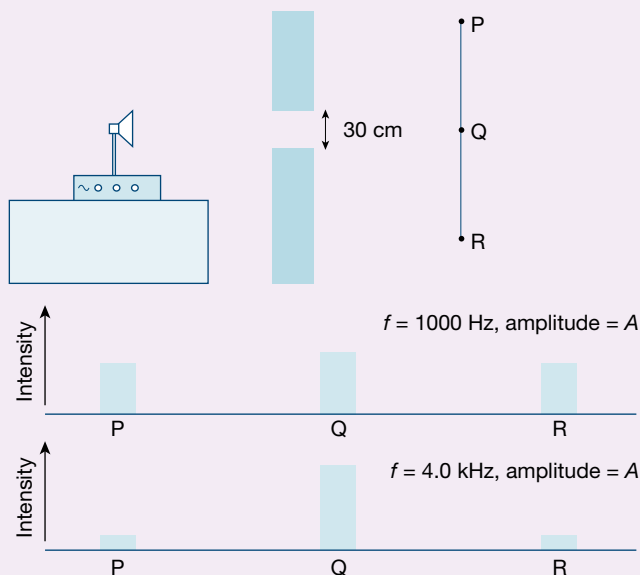


15.3 questions

Diffraction of sound

The following information relates to questions 1–7.

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 $f = 1000 \text{ Hz}$, and amplitude = A .

- 1 Calculate the wavelength of the sound produced.
- 2 Explain why noticeable sound intensity levels can be detected at points P and R.

The frequency of the signal generator is now set at $f = 4.0 \text{ kHz}$, amplitude = A .

- 3 What is the wavelength of this sound?
- 4 Explain why the intensity levels at points P and R are significantly less at this frequency than at 1000 Hz.
- 5 Why is the intensity of sound at point Q higher at this frequency?
- 6 With the frequency still set at 4.0 kHz, the amplitude of the sound is increased. Which of the following alternatives best describes the reason for an increased intensity being detected at points P, Q and R?
A Increasing the amplitude increases the kinetic energy of the waves, causing them to diffract more.

B A larger amplitude means more sound energy is being produced, so more will arrive at these points, resulting in an increase in intensity.

C Increasing the amplitude of the sound will also increase the wavelength, causing greater diffraction.

- 7 If the width of the opening is increased while the frequency is held constant, which of the following occurs?

A Only the higher frequencies are diffracted more than before.

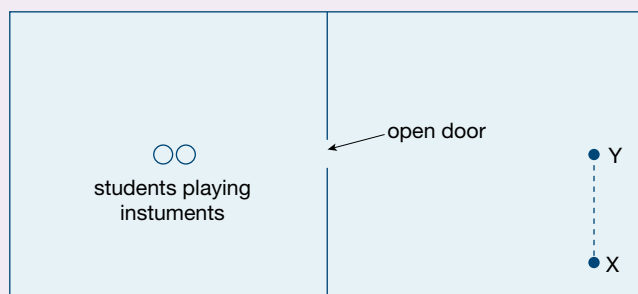
B Only the lower frequencies are diffracted more than before.

C All frequencies are diffracted more than before.

D All frequencies undergo less diffraction than before.

- 8 Explain why sonar equipment uses very high-frequency sound waves.

The following information relates to questions 9 and 10. The diagram depicts the floor plan of a music room where two students are practising. One plays a double bass while the other plays a violin. The music teacher, seated at point X, is able to listen to the students through an open door. Both students play continuously with equal loudness.



- 9 Which alternative best describes what the teacher hears?

A The teacher will hear the double bass more clearly than the violin.

B The teacher will hear the violin more clearly than the double bass.

C The teacher will hear both instruments with equal clarity.

- 10 If the teacher stands at point Y, which instrument will she hear more clearly? Justify your answer.

15.4 Amplitude, intensity and the decibel scale

Sounds are produced by vibrating objects. If more energy is supplied, the vibration will be larger and will last for a greater time and over a longer distance as the sound travels away from the source. An increase in the amplitude of the sound wave corresponds to an increase in the volume or loudness of the sound, but the frequency and speed of the sound remain independent of change in volume and amplitude.

It would be useful to revise the mathematics of logarithms and indices before commencing this section.



The **LOUDNESS** of a sound depends on the amplitude of the sound wave.

Loudness, amplitude and intensity

How a sound is heard depends on a number of factors. One person may hear a sound as louder or softer than another. This is referred to as the *perceived loudness* of the sound, and depends on the particular response of an individual's ear to each given frequency, as human hearing is not equally sensitive to all frequencies. Loudness is therefore partly a subjective judgement. A sound that one person can hear easily may be difficult to hear for another.

It is not easy to explain the relationship between the loudness of a sound and the sound wave's amplitude. Like amplitude, the loudness of a sound will depend on the energy that the sound wave is transferring.

The rate at which the energy is being carried by the sound wave through a given area is the *intensity* of the sound. Intensity is measured in watts per square metre (W m^{-2}). The human ear can detect sounds with an intensity as low as $10^{-12} \text{ W m}^{-2}$ (one-millionth of a millionth of a watt per square metre!). Sounds above 1 W m^{-2} will cause pain and may damage the ear.



The **INTENSITY** (I) of a sound wave is the rate at which the wave is carrying energy away from the sound source through a given area. It is the sound energy passing through a unit area each second.

Change in intensity with distance from source

A stone thrown into a pond produces a series of spreading ripples, and as these ripples move away from the point where the stone landed, their amplitude decreases. The energy of the original source is spread over a greater area. As the listener moves away from a sound source, a similar thing happens. The apparent loudness decreases and the sound becomes increasingly difficult to hear. Cupping your hands around your mouth, or using a simple megaphone, will limit the spread of the sound and allow it to be heard at a greater distance.

From a point source of sound, the sound will spread out radially: that is, both vertically and horizontally. When the distance from the source is doubled, the energy carried by the sound is spread over double the distance in all directions, or four times the original area. This means that the intensity will be reduced to a quarter of that at the original distance. At three times the original distance from the source, the intensity will be reduced to one-

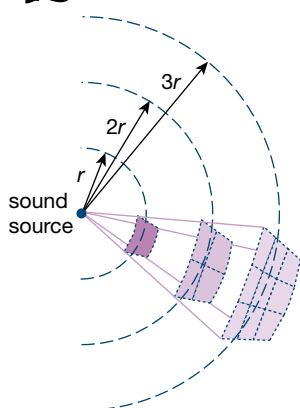


Figure 15.15 As it moves away from the source, a sound wave spreads radially. The same energy must spread over an area increasing with the square of the distance from the source, resulting in the intensity decreasing by the same ratio.

ninth of that at the starting point. This is referred to as an *inverse square law*. If the source is considered to be a point source, then the sound intensity, I , will decrease with the square of the distance, r , from the source:

$$I \propto \frac{1}{r^2}$$

The decrease in intensity follows this rule only if there is no other loss of energy within the medium. The decrease in sound with distance can often be much higher when the material through which the sound is travelling absorbs the energy of the wave. Wind and temperature differences in the medium can also affect how the intensity changes with distance.

Worked example 15.4A

An ambulance started its siren after leaving the scene of an accident, when it was 200 m away. The sound intensity heard at this distance was $8.0 \times 10^{-6} \text{ W m}^{-2}$. After a short time, the intensity of the siren's sound had fallen to $2.0 \times 10^{-6} \text{ W m}^{-2}$. Assuming the volume had not changed, how far was the ambulance from the accident scene at that time?

Solution

We know that the intensity (I) is inversely proportional to the square of the distance (d). The original intensity (I_1) was $8.0 \times 10^{-6} \text{ W m}^{-2}$ at 200 m, and the final intensity (I_2) was $2.0 \times 10^{-6} \text{ W m}^{-2}$, so we know that $I_1/I_2 = 4$.

$$I_1 \propto \frac{1}{d_1^2} \text{ and } I_2 \propto \frac{1}{d_2^2}$$

$$\text{therefore } \frac{I_1}{I_2} \propto \left(\frac{d_2}{d_1}\right)^2$$

$$\text{Since } \frac{I_1}{I_2} = 4, \text{ we have } 4 = \left(\frac{d_2}{d_1}\right)^2$$

That is, $2 = d_2/d_1$ or $d_2 = 2d_1 = 2 \times 200 \text{ m}$. The ambulance is therefore 400 m away.

The solution may have been obvious in this example, but the method applies equally to more complex examples.

Physics file

An important distinction is made here between intensity and sound level. A sound's *intensity* (I) is measured in watts per square metre (W m^{-2}), while its *level* (L) is measured in decibels (dB).

Sound level and the decibel scale

When a sound wave reaches your ear, changes in pressure between rarefaction and compression cause the eardrum to vibrate. There is a great difference between the maximum pressure at compressions (the *pressure amplitude*) of a quiet sound and that of a loud one. The maximum air pressure amplitude produced by a jet aircraft taking off or a rock band can be one million times that of a soft whisper or a pin dropping. This does not mean that loud sounds are heard, or *perceived*, to be many thousands of times louder than quieter sounds. The human ear compares sounds in terms of the *ratio* of their pressure amplitudes rather than the *difference* in amplitudes. Thus, doubling the sound of a quiet whisper is very noticeable, while whispering in a crowd is barely noticeable.

To enable us to compare sounds in the way human hearing does, the *decibel scale* was devised. This scale simulates the natural response of the human ear by comparing the intensity of a sound wave with a standard value. The basic unit of the decibel scale is the *bel*. Because this is too large for use in the normal range of sound levels, the *decibel* (dB) is commonly used: $1 \text{ dB} = 0.1 \text{ bel}$.

Because the ear's ability to hear different frequencies varies, the sound levels on this scale are compared at a frequency of 1 kHz. The scale is logarithmic:

$$L = 10 \log \frac{I}{I_0}$$

where L = sound level (dB)
 I = intensity of the sound source (W m^{-2})
 I_0 = intensity of a reference source (W m^{-2}).



The **DECIBEL SCALE** is based on the measurement of the ratio of the intensity of the sound to a standard intensity at 1 kHz.

The logarithmic nature of this scale makes it easier to compare the huge range of intensities we normally hear. The standard reference intensity is usually accepted as the minimum intensity which can be heard by the average person. This intensity (I_0), $10^{-12} \text{ W m}^{-2}$, is referred to as the *threshold of hearing*. At a frequency of 1 kHz, a sound of intensity $10^{-12} \text{ W m}^{-2}$ is 0 dB; 0 dB can be heard—but only just!

Based on this scale:

- Increasing the intensity of a sound by a factor of 10 (e.g. from 10^{-6} W m^{-2} to 10^{-5} W m^{-2}) corresponds to an increase of 10 dB in the sound level. Increasing the intensity 100 times corresponds to a 20 dB increase, and so on. Conversely, increasing a sound level by 20 dB means a $10 \times 10 = 100$ times increase in intensity.
- Increases in sound level in equal steps (e.g. in 20 dB steps: from 20 dB to 40 dB and then to 60 dB) are perceived by the human ear as approximately equal increases in loudness.
- The normal human ear can distinguish differences in sound levels down to about 1 dB under ideal conditions.

An increase in sound level of 3 dB corresponds to a doubling of the sound intensity. (This can be useful in quickly comparing intensities.)

Table 15.2 Sound intensity levels for some common sounds

Sound source	Sound intensity (W m^{-2})	Sound level (dB)
Threshold of human hearing	10^{-12}	0
Rustling leaves	10^{-11}	10
Soft whisper	10^{-10}	20
Normal conversation	10^{-6}	60
Street traffic	10^{-5}	70
Car alarm	10^{-2}	100
Indoor rock concert, front row	1	120
Threshold of pain	1	120
Jet aircraft at 30 m	100	140
Space shuttle launch at 50 m	108	200

Physics file

Sound level meters and sensors are designed to compare sounds in a similar manner to the human ear. Middle frequencies (300–3000 Hz) in the audible range are better detected than extremes of the range.

An audio engineer's high-quality sound-measuring equipment allows for different frequencies and adjusts the readings accordingly: the scale is *frequency weighted*. It is referred to as the *A-scale* and is measured in units of the *A-weighted decibel* or dB(A). This is the most widely used scale, since it best matches human hearing. For the purposes of this study, it is reasonable to assume that dB readings represent both scales.



Figure 15.16 The high pressures and loud sounds of wartime explosions far exceed the levels at which the human ear was meant to perform. As a result, ruptured eardrums and deafness have been common battlefield injuries since the introduction of the first artillery.

Worked example 15.4B

The sound level of a particular sound increases from 40 to 43 dB. What is the ratio of the final sound intensity to the original sound intensity?

Solution

This can be solved easily by using the fact that every increase in sound level of 3 dB corresponds to a doubling of the intensity. The increase from 40 to 43 dB is 3 dB, so the final sound intensity is double the original, and the ratio is 2.

Worked example 15.4C

A sound level of 120 dB is regarded as the threshold of pain for the average listener.

- a What is the corresponding sound intensity?
- b If a listener is exposed to a sound level of 120 dB at a distance of 10 m from the source, to what value would the level fall if they moved a further 10 m away?

Solution

- a We know that $L = 120$ dB and $L = 10 \log \frac{I}{I_0}$.

$$\text{Since } I_0 = 10^{-12} \text{ W m}^{-2}, \quad 120 = 10 \log \frac{I}{10^{-12}}$$

$$\text{so } 12 = \log \frac{I}{10^{-12}}$$

$$\text{i.e. } 10^{12} = \frac{I}{10^{-12}}$$

$$\text{so } I = 10^{12} \times 10^{-12} = 1 \text{ W m}^{-2}$$

- b The distance would be doubled, and since $I_1 \propto \frac{1}{d^2}$, where I_1 = intensity at 10 m, at the new position $I_2 \propto \frac{1}{(2d)^2} \propto \frac{1}{4d^2} \propto \frac{1}{4} \times \frac{1}{d^2}$.

Since $I_1 \propto \frac{1}{d^2}$, $I_2 = \frac{1}{4} \times I_1$, i.e. the intensity is $\frac{1}{4}$ of the original, so $I_2 = \frac{1}{4}$ or 0.25 W m^{-2} .

$$\begin{aligned} \text{As a sound level: } L &= 10 \log \frac{0.25}{10^{-12}} \\ &= 10 \log (2.5 \times 10^{11}) \\ &= 10 (\log 10^{11} + \log 2.5) = 10(11 + 0.398) = 10 \times 11.398 \end{aligned}$$

The new sound level would be about 114 dB.



Figure 15.17 Despite advances in amplifier design, the 1970s band Black Sabbath still retains the dubious honour of producing the loudest sound at a rock concert—a sustained level of over 140 dB.

Is sonar killing whales?

The world's oceans, already stressed by pollution, global warming and the effects of increased UV exposure, are not nearly as silent as we often imagine them. The sounds of human oil exploration can often exceed 180 dB—compare that with the sound level causing pain in humans! Whales too find sound levels greater than 120 dB uncomfortable: they turn away, but sounds travel vast distances in the oceans. In 2000, an apparent connection was established between whale beachings and sonar or seismic activity. Since then, scientists and environmentalists have been paying close attention to the effects human noise is having on whales.

In 2000, 16 whales were beached during a US Navy exercise in the Bahamas. Clear evidence of bleeding and tissue damage around the ears of the dead whales led to a halt in US Navy sonar experiments and a seismic survey. Following a second incident in the Canary Islands, when 15 beaked whales were beached, a US Federal Court granted a preliminary injunction blocking worldwide use of low-frequency active sonar. But this technology is only part of the problem.

In an empty ocean, simple water waves cause sound levels of around 80 dB. The loudest sounds used for seismic investigations generally have sound levels of up to 225 dB—every 10 seconds.

Oil exploration in the Otway Basin off western Victoria and South Australia uses sounds of up to 240 dB, which are still at levels of around 150 dB at 10 km from the source. Current rules by Environment Australia only call for seismic work to be stopped if whales are seen within 3 km. Strong evidence is emerging that these sounds may injure and disorient whales, who rely heavily on sound to navigate and communicate over long distances. Scientists also believe that excessive noise causes whales to change their migration patterns and that it disrupts their songs used for attracting mates, potentially leading to a long-term population decline.

Sightings suggest that whales are returning to an area close to where the scientific tests are taking place. Are the whales growing accustomed to the sound or so hungry that they have no choice?



Figure 15.18 While humans find sound levels above 120 dB uncomfortable or painful, whales are regularly enduring sound levels in excess of 180 dB during seismic investigation. Beached whales have been found with evidence of bleeding and tissue damage around the ears.



15.4 summary

Amplitude, intensity and the decibel scale

- The loudness of any sound is related to the amplitude and frequency of the sound wave.
- The perceived loudness of a sound depends on the response of an individual's hearing to a particular frequency.
- The intensity (I) of a sound wave is the rate at which the wave is carrying energy through an area of 1 square metre; it is measured in W m^{-2} . The human ear can detect sounds as soft as $10^{-12} \text{ W m}^{-2}$ and up to 1 W m^{-2} without pain.
- The intensity of sound decreases with the square of the distance from a point source, i.e. $I \propto \frac{1}{r^2}$.
- Sound intensity is a measure of power per square metre (W m^{-2}). Sound level refers to measurements in decibels (dB).
- The decibel scale simulates the natural response of the human ear by comparing the intensity of a sound wave with a standard value.
- The sound level is defined by the relationship:

$$L \text{ (in dB)} = 10 \log \frac{I}{I_0}$$
 where I is the intensity of the sound (W m^{-2}) and I_0 is the intensity of a reference source (W m^{-2}). The reference source is usually adopted as the threshold of hearing at 1 kHz, which is $10^{-12} \text{ W m}^{-2}$.

- Increasing a sound's intensity by a factor of 10 increases its sound level by 10 dB. An increase in sound level of about 3 dB corresponds to a doubling of the intensity of the sound. Similarly, a decrease in sound level of about 3 dB corresponds to a halving in the intensity of the sound.

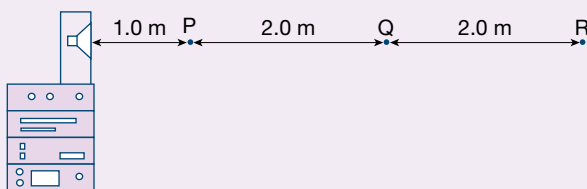


15.4 questions

Amplitude, intensity and the decibel scale

- Turning up the volume control on a radio produces sound with which of the following characteristics?
 - higher frequency
 - higher wavelength
 - higher speed
 - higher intensity
- Two people are listening to the same music at the same distance. They disagree on its loudness. Explain how this could happen.
- The intensity of the sound received at a particular point depends on:
 - the distance from the source only
 - the power output of the source only
 - both the power output of the source and the distance from the source
 - the power output of the source, distance from the source and properties of the medium
 - the power output of the source, distance from the source, properties of the medium and the prevailing conditions (temperature, pressure etc.).

The following information relates to questions 4–10.



A loudspeaker is connected to a signal generator and amplifier. It is set up in an acoustic laboratory as shown in the diagram. Assume that the loudspeaker acts as a point source and that no sound is reflected from the surrounding surfaces. The speed of sound in the laboratory is 340 m s^{-1} . The frequency produced by the signal generator is set at 1.5 kHz.

- A version of the decibel scale adjusted for variations in our perception of loudness at different frequencies is referred to as the A-scale, with the unit dB(A).

- What is the wavelength of the sound at:
 - point P?
 - point Q?

A student places a sound intensity meter at point P and obtains a reading of $4.0 \times 10^{-2} \text{ W m}^{-2}$.

- What is the intensity of the sound at:
 - point Q?
 - point R?

The output of the signal generator is changed so that the power output of the loudspeaker is now 1.0 W.

- What is the new sound intensity at:
 - point Q?
 - point R?
- How much energy does this loudspeaker produce during a 10-second interval?
- Calculate the sound levels in dB for a person standing at:
 - point P
 - point Q
 - point R.
- A neighbour asks for the sound to be turned down. The power output of the loudspeaker is adjusted so that the sound intensity level at point P is 100 dB.
 - What is the sound intensity at P?
 - Calculate the intensity at Q.
 - Calculate the sound level at Q.
- By installing insulation, the sound level of a city office is reduced by 12 dB. If the final intensity is $5.0 \times 10^{-10} \text{ W m}^{-2}$, estimate the sound intensity before the insulation was installed.
- The sound from a pneumatic drill is 1.0 mW m^{-2} at a distance of 1.0 m. Assume that the sound spreads out equally in all directions.
 - Calculate the intensity at a distance of 10 m.
 - What is the sound level at a distance of 10 m?

15.5 Frequency, perceived loudness and the phon

If two sounds have different frequencies, a listener might hear one but not the other, even if the sounds have the same intensity. The differing response of the human ear is referred to as the perceived loudness of a sound.

The ability of a loudspeaker to accurately produce the full range of frequencies represented in the amplified sounds is a major design characteristic of any sound system. A person may hear one frequency better than another because of a better personal response to particular frequencies, but overall a speaker system should produce a sound with a response independent of the listener.

Frequency and pitch

The frequency of light determines its colour. In a similar way, the frequency of a sound wave determines the *pitch* of the sound. While frequency is the major factor in determining the particular pitch of a sound, there is a definite distinction between 'frequency' and 'pitch'.

Frequency is a particular physical characteristic of a sound wave. It is not affected by the listener or by the location. There does not even need to be a listener. The pitch, however, is a particular physiological or psychological sensation—related to the frequency—which depends entirely on the listener. A tone-deaf person might not be able to tell one sound from another, but different frequencies would still be there.

The frequencies which the average young person can hear range from about 20 Hz to 20 kHz. Sounds outside this range are referred to as *infrasonic* if they are lower than 20 Hz and *ultrasonic* if they are higher than 20 kHz. Human voices fall within the range of 85–1100 Hz.

As people age, the range of frequencies they can hear begins to decrease, particularly in the higher frequency range. By the age of 65, many people cannot hear sounds beyond 5000 Hz. As most common sounds lie below this frequency, the reduced range does not always create problems.

Continuous loud sounds can also lead to premature loss of hearing range. Regardless of the source and range of the noise, loud sounds with frequencies between 2000 and 5000 Hz can cause hearing loss. Industrial guidelines for exposure to noise recommend exposure of no more than 6 hours a day to sounds of 85 dB.

The frequency response of human hearing

Regardless of age or environmental effects on hearing, the human ear does not respond equally well to all frequencies. Generally, the ear responds best to sounds in the range of 300–3000 Hz. Lower or higher frequency sounds need to have a significantly greater sound level to be heard at the same perceived loudness as sounds in this middle range.

For a particular listener to hear sounds of different frequencies at the same loudness, quite different intensities may be required. And a particular sound may appear to change quite markedly in loudness when the frequency is changed, even though there is no change in amplitude. A graph of the frequency response for a typical ear gives an indication of the relative perceived loudness of differing frequencies.

Physics file

Many animals can hear a significantly greater range of frequencies than humans. Animals that rely on sound for navigation produce high-frequency directional sounds that lie well outside the range of human hearing.

Table 15.3 Decline in frequency response with age

Age (years)	Highest audible frequency (Hz)
3–20	20 000
35	16 000
45	12 000
55	8 000
65	5 000

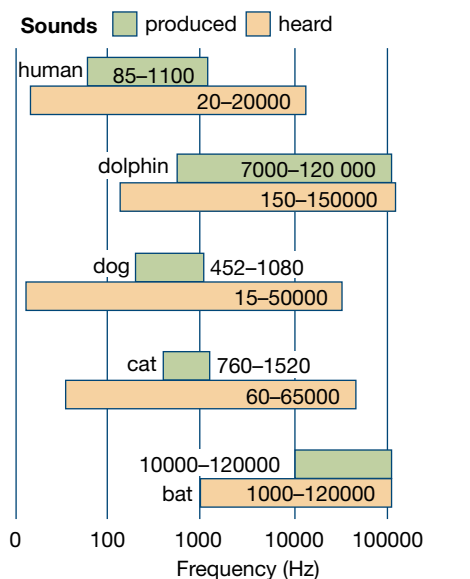


Figure 15.19 The frequencies of sound that different animals can produce and detect compared with the human range.

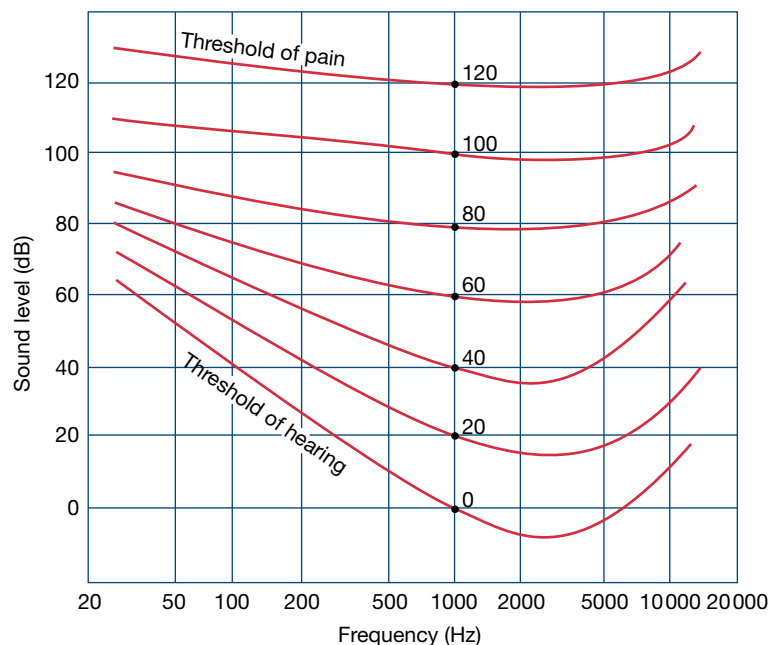


Figure 15.20 Graph of the sensitivity of the human ear. The curves indicate the sound levels that are perceived as equally loud. The number labelling each curve represents the loudness level in phons that is numerically equal to the level at 1000 Hz. For example, a sound of frequency 50 Hz must have a sound level of 80 dB to be perceived to be as loud as a 1000 Hz sound would at 60 dB or a 10 000 Hz sound would at 70 dB. To be perceived as equally loud, higher and lower frequencies generally need to have higher sound levels than middle frequencies.

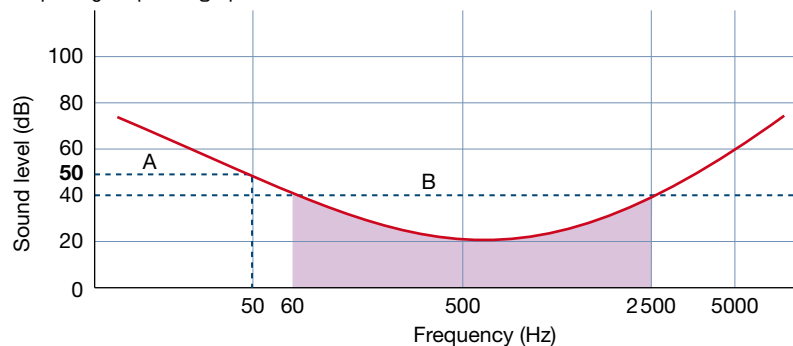


PRACTICAL ACTIVITY 48

Pitch, loudness and quality

Worked example 15.5A

A loudspeaker emits only one frequency at a time. The output level is gradually increased and decreased and a listener notes at what sound level he can hear the sound. The test is repeated for a range of frequencies and the results are plotted to give the following frequency response graph.



- What sound level at 50 Hz gives the same perceived loudness as a level of 20 dB at 500 Hz?
- If the sound level produced by the loudspeaker was held at a constant 40 dB while the frequency was varied, what range of frequencies would be heard?

Solution

Both questions can be answered directly from the graph.

- Reading off the graph (as indicated by line A), the required sound level is 50 dB.
- Line B represents a sound level of 40 dB. It intersects the frequency response curve at 60 Hz and again at 2500 Hz. This is the range which the listener can hear. Outside this range a sound level of 40 dB is not sufficient for the person to hear the sounds.

Infrasonics and ultrasonics

Normal human ears can detect sounds with frequencies between 20 Hz and 20 kHz. Sounds below 20 Hz are called *infrasound* and those above 20 kHz are called *ultrasound*.

Infrasound is virtually non-directional: it is difficult, if not impossible, to determine its source. Geologists use infrasound in seismographic surveying, where the vibrations generated by underground explosions can give an indication of the geological structure. Modern movie theatres often have large banks of speakers producing large-amplitude, low-frequency sound waves to produce a sensation known as 'Sensurround'. To give the movie-goer a more realistic feeling of, say, an earthquake or a bomb blast, subsonic sounds of around 3 Hz can be played via these speakers. The body detects these sounds as vibrations.

Ultrasound has many more uses. Many animals, such as bats and dolphins, use ultrasonic pulses to locate objects around them and to communicate. The highly directional nature of the reflected ultrasonic sounds allows the accurate location of even very small objects. Many marine applications have been found for ultrasonic 'sonar' (**s**ound **n**avigation and **r**anging) since its invention during World War II. It can be used by warships to locate submarines, and for general shipping to map the sea floor, locate underwater obstacles, detect changes in water temperature and find shoals of fish.

In industry, ultrasound can be used to detect otherwise invisible flaws in metal structures. In medicine, it is used to obtain pictures of the internal structures of the body, particularly during pregnancy when ultrasound is considered less harmful than X-rays. Many schools now also regularly use ultrasonic sensors to investigate motion.



Figure 15.21 Ultrasound can be used to give a clear picture of the developing foetus. It allows early detection of potentially serious birth defects and, if parents wish, of the sex of the unborn child. The head of the foetus can be seen at the right side of this picture.



15.5 summary

Frequency, perceived loudness and the phon

- The human ear perceives different frequencies of the same intensity as having different sound levels. This is called the perceived loudness of sounds.
- The pitch of a sound is related to its frequency, but is dependent on the listener.
- Normal human hearing covers a frequency range of 20–20 000 Hz. Sounds outside this range are called subsonic (less than 20 Hz) or ultrasonic (above 20 000 Hz). Human hearing responds best to sounds between 300 and 3000 Hz.
- Frequency response curves, or phon curves, allow comparisons of the perceived loudness of different frequencies based on a standard sound of 1000 Hz.



15.5 questions

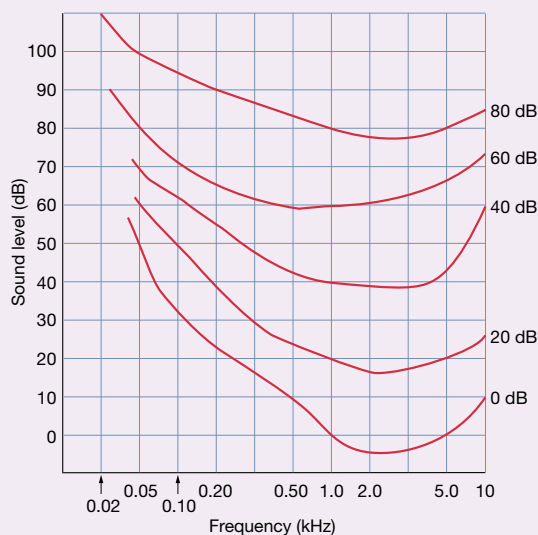
Frequency, perceived loudness and the phon

- 1 Explain the difference between pitch and frequency.
- 2 Which of the following would a healthy young person be able to detect?
 - A All frequencies
 - B Frequencies greater than 20 kHz
 - C Frequencies less than 20 kHz
 - D Frequencies from 20 Hz to 20 kHz

- 3 Which of the following is true of the loss of hearing as a person gets older?
- A It is uniform across the whole range of audible frequencies.
 - B It occurs only for frequencies between about 3 and 8 kHz.
 - C It is mainly restricted to frequencies above 10 kHz.
- 4 Which of the following explains why the phenomenon of 'perceived loudness' occurs?
- A The human ear is more sensitive to some frequencies than to others.
 - B Higher frequencies carry more energy than lower frequencies.
 - C Our hearing deteriorates as we get older.

The following information relates to questions 5–8.

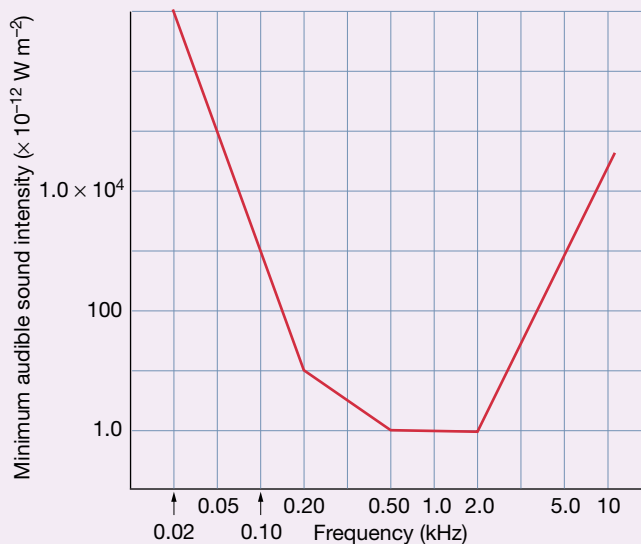
A hearing test was conducted on an airforce recruit to see how she responded to sounds of different intensities and frequencies. The following graph shows the results. Each curved line connects the sound levels at which she perceives the same loudness as she would hear for a frequency of 1.0 kHz at that level, i.e. 1000 Hz is the reference frequency for loudness.



- 5 At what frequency will this recruit perceive a 20 dB sound to have the same loudness as a sound of:
- a 40 dB?
 - b 50 dB?
 - c 60 dB?
- 6 What intensity level must a 200 Hz sound have for her to perceive it as having the same loudness as a 80 dB sound?

- 7 For sounds of a level of 60 dB, what is the range of frequencies for which she judges the perceived loudness to be approximately constant?
- 8 While the recruit was being subjected to a test sound of frequency 1.0 kHz and sound level 60 dB, an 80 dB foghorn was activated. The recruit noticed that both sounds seemed equally loud. Which of the following is the best estimate of the frequency of the foghorn?
- A 1.0 kHz
 - B 2.0 kHz
 - C 50 Hz
 - D 500 Hz

The following information relates to questions 9 and 10. In an experiment in an acoustic laboratory, the minimum audible sound intensity produced at a distance of 1.0 m from a loudspeaker was recorded for various frequencies. The following graph shows the minimum audible sound as a function of frequency for this loudspeaker.



- 9 a What is the power output of this loudspeaker for a minimum audible intensity of $1.0 \times 10^{-8} \text{ W m}^{-2}$?
- b What is the minimum audible sound level for a frequency of 5.0 kHz?
- 10 Calculate the value of the ratio in decibels of the minimum audible sound intensity at 10 kHz to that at 2.0 kHz.

15.6 Making sound: Strings and air columns

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.

Superposition

Imagine two transverse waves travelling in opposite directions along a string, as shown in Figure 15.22a. 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 15.22b, 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.

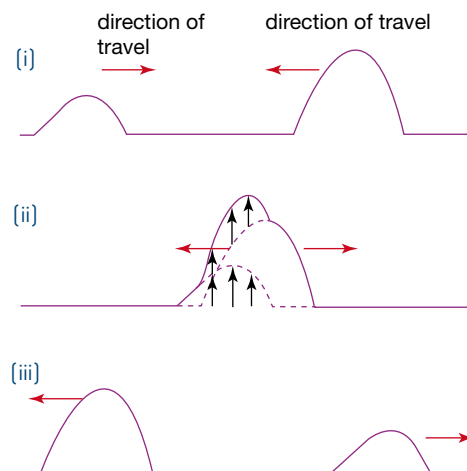


When two or more longitudinal or transverse waves meet, the resulting displacement at each point will be the vector sum of the displacements of the component waves. This is the principle of **SUPERPOSITION**.



Figure 15.23 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.

(a) Constructive superposition of waves



(b) Destructive superposition of waves

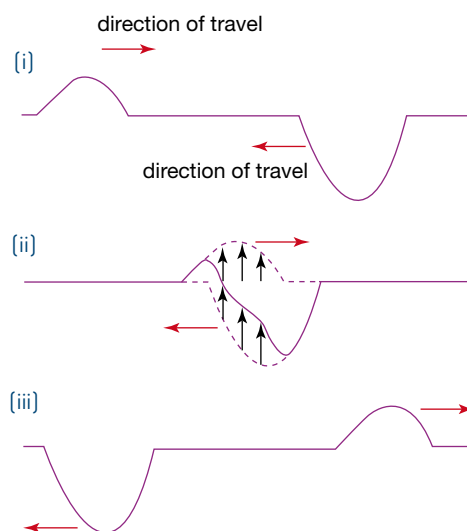


Figure 15.22 Superposition of waves in a string. After the interaction, the pulses continue unaltered: they do not permanently affect each other.

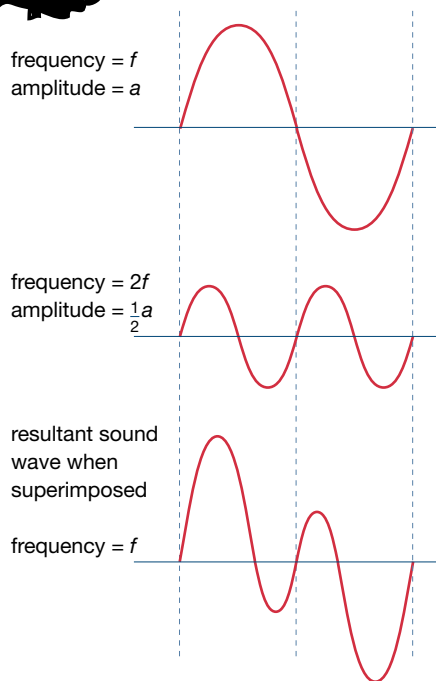


Figure 15.24 Two sound waves, one twice the frequency of the other, produce a complex wave of varying amplitude when they are superimposed.

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. 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 15.25).

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:

$$f_{\text{beat}} = |f_1 - f_2|$$

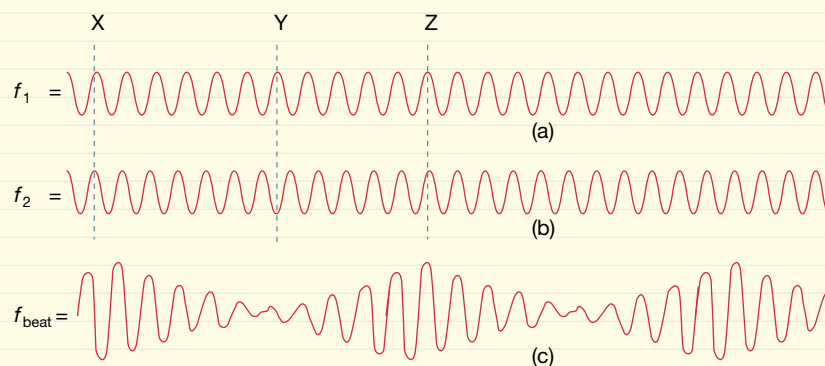


Figure 15.25 A beat (f_{beat}) produced by two sound waves f_1 and f_2 of equal amplitude but slightly different frequency. At time X the waves are in phase, producing a larger amplitude. A short time later, at Y, they are out of phase, resulting in a smaller amplitude, and at Z they are in phase once 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.

The effects of superposition are around us all the time. The ripples in the pond of Figure 15.23 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 15.24). 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.

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 15.26.

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.

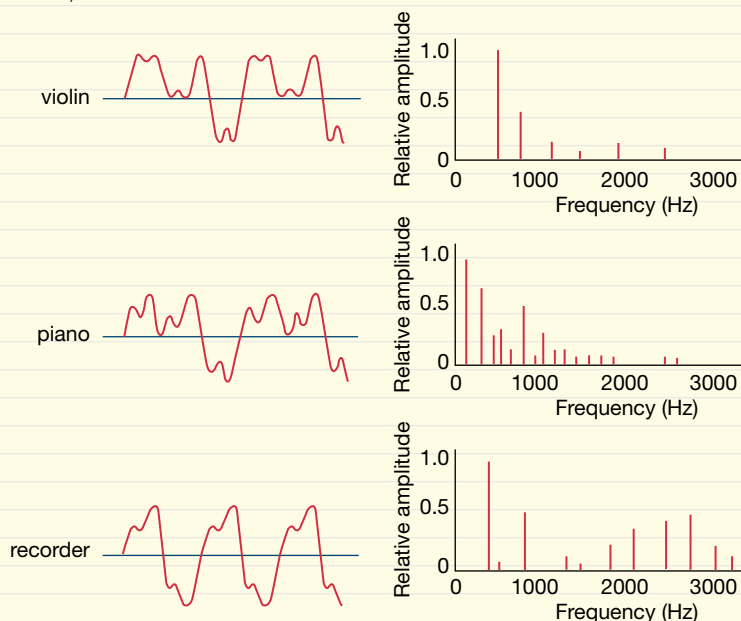


Figure 15.26 The complex waveforms produced by the violin, piano and recorder result from the superposition of a range of frequencies with differing energies and durations.

Resonance

You have probably heard about singers who can break glass by singing particularly high notes. Figure 15.27 shows a glass being broken in much the same way. The 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.

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’ve established would *decrease* the total energy.) This is an example of a *forced*



Figure 15.27 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.

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.

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.

Figure 15.28 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.



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.

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 a 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 a 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 15.30. It has only two nodes, and so the length of the string, L , will correspond to $\lambda/2$.

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 15.31 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

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 15.29. 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.

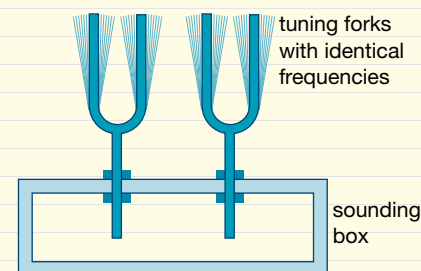


Figure 15.29 When one tuning fork is struck, the second identical fork will vibrate as a result of resonance.

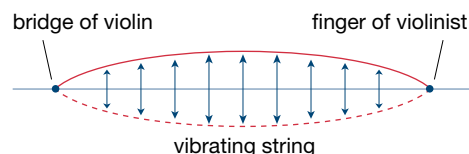


Figure 15.30 The simplest mode of vibration of a bowed violin string is an example of a transverse standing wave.

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 15.32. 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 15.31 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.

T = the period of each wave
(i.e. the time to go through one cycle or to travel one wavelength)

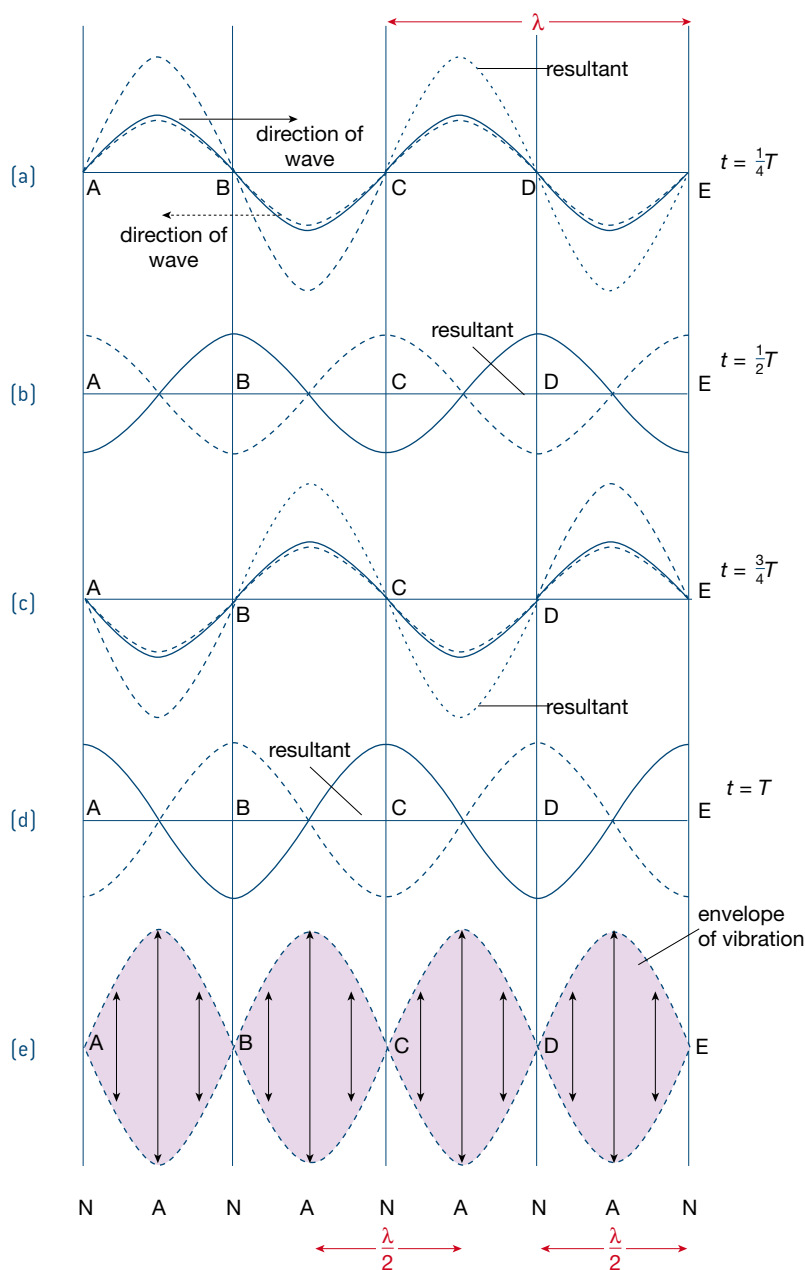
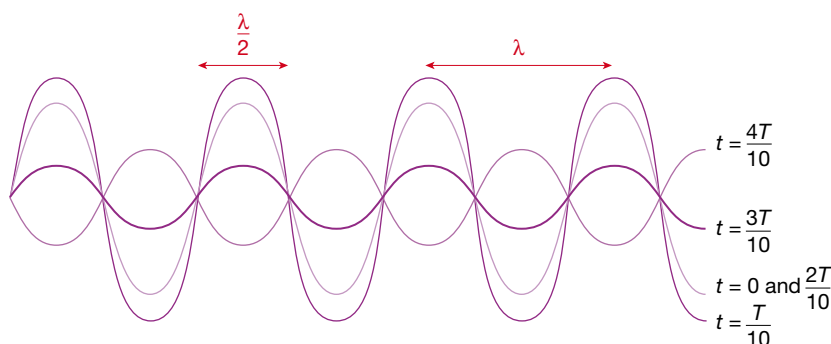


Figure 15.32 Within the wave envelope, a transverse standing wave may appear to have a variety of amplitudes, from $-2A$ to $+2A$. However, the nodes remain stationary.



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 harmonics are referred to as *overtone*s. Figure 15.33 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, 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, $L = \lambda/2$. The second and third harmonics satisfy the conditions $L = 2\lambda/2$ and $L = 3\lambda/2$ respectively.

In general, $L = \frac{n}{2} \times \lambda_n$ or:

$$\lambda_n = \frac{2L}{n}$$

where λ = wavelength (m)

L = length of the string (m)

n = number of the harmonic = 1, 2, 3, ...

Using the wave equation, $v = \phi\lambda$, the relationship between frequency, velocity and string length can be established:

$$f = \frac{nv}{2L}$$

where f = frequency of the harmonic (Hz)

v = velocity of the component waves (m s^{-1}).

Worked example 15.6A

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 15.6B

A violin string has a length of 22 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, $n = 1$, so $L = \lambda/2$ and thus $\lambda = 2L = 2 \times 0.22 = 0.44$ m.

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.

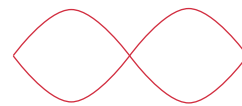
The seventh harmonic of a violin string produces the high-pitched squeal commonly made by beginners. The solution is to bow the strings at the position of the first node for the seventh harmonic: then the standing wave of this harmonic cannot form and the squeal is not heard. The standing waves corresponding 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.)

first harmonic
(fundamental frequency)



$$\lambda_1 = 2L \quad f_1 = \frac{v}{\lambda_1} = \frac{v}{2L}$$

second harmonic
(first overtone)



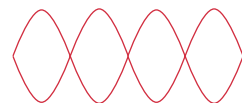
$$\lambda_2 = L \quad f_2 = \frac{v}{\lambda_2} = \frac{v}{L} = 2f_1$$

third harmonic
(second overtone)



$$\lambda_3 = \frac{2L}{3} \quad f_3 = \frac{v}{\lambda_3} = \frac{3v}{2L} = 3f_1$$

fourth harmonic
(third overtone)



$$\lambda_4 = \frac{L}{2} \quad f_4 = \frac{v}{\lambda_4} = \frac{2v}{L} = 4f_1$$

Figure 15.33 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.

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 15.34 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 15.35 The flute is a typical example of a pipe open at both ends where an air column can be made to vibrate.

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

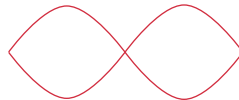
At the open end of a pipe, sound waves are reflected. When a compression or rarefaction reflects from an open end, it does so with a phase change of $\lambda/2$.

first harmonic
(fundamental frequency)



$$\lambda_1 = 2L \quad f_1 = \frac{v}{\lambda_1} = \frac{v}{2L}$$

second harmonic
(first overtone)



$$\lambda_2 = L \quad f_2 = \frac{v}{\lambda_2} = \frac{v}{L} = 2f_1$$

third harmonic
(second overtone)



$$\lambda_3 = \frac{2L}{3} \quad f_3 = \frac{v}{\lambda_3} = \frac{3v}{2L} = 3f_1$$

fourth harmonic
(third overtone)



$$\lambda_4 = \frac{L}{2} \quad f_4 = \frac{v}{\lambda_4} = \frac{2v}{L} = 4f_1$$

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:

$$\lambda_n = \frac{2L}{n}$$

where λ = wavelength (m)

L = effective length of the air column (m)

n = 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:

$$f_n = \frac{nv}{2L}$$

where f = frequency of the harmonic (Hz)

v = 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 15.34. 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 ($f_3 = 3f_1 = 3/2 f_2$), and so on.

Typical open-ended pipes include flutes, oboes and similar instruments and the muffler in a car exhaust system.

Worked example 15.6C

A particular flute has an effective length of 35 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} .

- What is the wavelength of the second harmonic?
- What will the frequency of the second harmonic be?

Solution

- We know the air column is 35 cm long, or 0.35 m. For the second harmonic, a drawing of the appropriate standing wave will show that:

$$\lambda = \frac{2L}{2} = 0.35 \text{ m}$$

- Using the wave equation and rearranging:

$$f = \frac{v}{\lambda} = \frac{350}{0.35} = 1000 \text{ Hz}$$

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 15.36). 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 $\frac{4}{3}$ the length of the air column; the next, $\frac{4}{5}$, and so on. In general:

$$\lambda_n = \frac{4L}{n}$$

and

$$f_n = \frac{nv}{4L}$$

where λ_n = wavelength of the harmonic (m)

L = effective length of the air column (m)

f_n = frequency of the harmonic (Hz)

v = velocity of the component waves (m s^{-1})

n = number of the harmonic = 1, 3, 5, ... (odd numbers only).



PRACTICAL ACTIVITY 49

Standing waves in air columns

Physics file

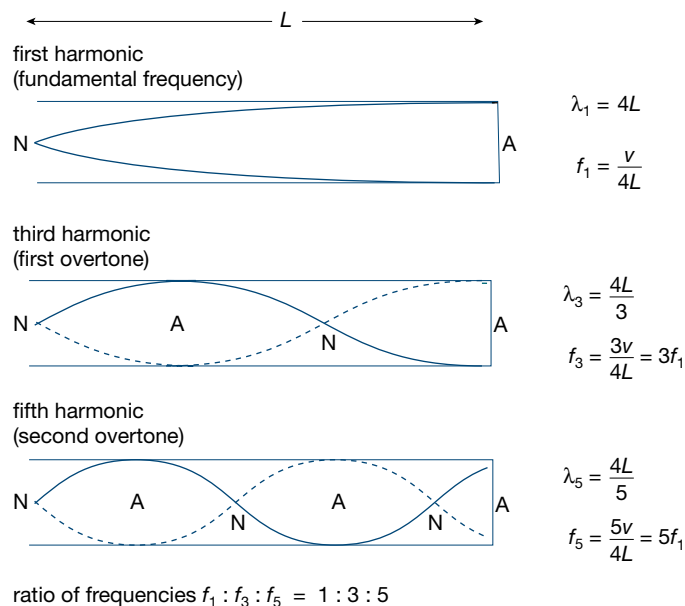
If the end of a pipe is flared (e.g. in a trumpet or saxophone) the situation is far more complex than for a plain end. As a wavefront leaves the relatively constant diameter of the main tube and enters the horn region of the pipe, it begins to spread and behave in a different manner. This maximises the amount of energy transmitted to the surrounding air. Research is still continuing on exactly how sound behaves in this region.



PRACTICAL ACTIVITY 50

Speed of sound by resonance tube

Figure 15.36 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 15.36 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 ($f_5 = 5f_1 = 5/3f_3$), 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 (see Figure 15.45, page 593), car engine manifolds and some organ pipes.

Worked example 15.60

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 3.0 cm long in an adult. Assume that the speed of sound is 340 m s^{-1} .

- What is the fundamental resonant frequency of the ear canal?
- What is the frequency of the next resonating frequency?
- Explain one reason why some frequencies are heard better than others.

Solution

- 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 15.36 shows that for a closed tube, the length will correspond to $L = \lambda/4$ and so:

$$\lambda = 4L = 4 \times 3 \text{ cm} = 12 \text{ cm or } 0.12 \text{ m}$$

Now, $f = \frac{v}{\lambda}$, and (from the question) $v = 340 \text{ m s}^{-1}$, so:

$$f = \frac{v}{\lambda} = \frac{340}{0.12} = 2833 \text{ Hz, or about } 2.8 \text{ kHz}$$

Compare this with the frequencies heard best by humans.

- As this is a closed tube, only odd-numbered harmonics will form, so the next harmonic will be the third: $f_3 = 3 \times 2833 = 8500 = 8.5 \times 10^3 \text{ Hz or } 8.5 \text{ kHz}$.
- 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.

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.

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



Figure 15.37 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.

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 $^{12}\sqrt{2}$, or about 1.06:1.

Table 15.4 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#, Bb		233
B		247
C (middle)	do	262
C#, Db		277
D	re	294
D#, Eb		311
E	mi	330
F	fa	349
F#, Gb		370
G	so	392
G#, Ab		415
A	la	440
A#, Bb		466
B	ti	494
C	do	523



15.6 summary

Making sound: strings and air columns

- 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 $f = \frac{nv}{2L}$. All harmonics ($n = 1, 2, 3, \dots$) may be present, and the ratio of frequencies is $f_1 : f_2 : f_3 : \dots = 1 : 2 : 3 : \dots$.

- 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.

$$\lambda_n = \frac{2L}{n} \text{ and } f_n = \frac{nv}{2L}$$

where n = number of the harmonic = 1, 2, 3, ...

All harmonics are possible, and the ratio of frequencies is $f_1 : f_2 : f_3 : \dots = 1 : 2 : 3 : \dots$.

- 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:

$$\lambda_n = \frac{4L}{n} \text{ and } f_n = \frac{nv}{4L}$$

where n = number of the harmonic = 1, 3, 5, ...

Only odd-numbered harmonics are possible, and the ratio of frequencies is $f_1 : f_3 : f_5 : \dots = 1 : 3 : 5 : \dots$.

- A rise in tone of one octave is equivalent to a doubling of the frequency.

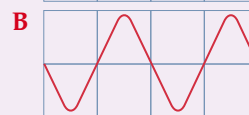
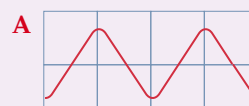


15.6 questions

Making sound: Strings and air columns

- State if each of the following statements concerning the interaction of two pulses is true or false.
 - The displacement of the resultant pulse is equal to the sum of the displacements of the individual pulses.
 - As the pulses pass through each other, the interaction permanently alters the characteristics of each pulse.
 - After the pulses have passed through each other, they will have the same characteristics as before the interaction.
 - The interaction will cause a loss in kinetic energy of each pulse, resulting in the pulses separating with reduced velocities.
- Why is it possible to shatter a glass by exposing it to sound of a certain frequency?
- How does an acoustic guitar amplify the sound it produces?
- Explain how a standing wave is produced.

- 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** $t = T/4$
- b** $t = T/2$
- c** $t = 3T/4$

- 6** At what point should a guitar string be plucked to:
- a** make its fundamental frequency most prominent?
 - b** make its second harmonic most prominent?
 - c** make its third harmonic most prominent?
- 7** The speed of a transverse wave in a metal string of 50 cm length, when subjected to a certain tension, was calculated to be 300 m s^{-1} . If this string were to be plucked, calculate the frequency of:
- a** the first harmonic
 - b** the second harmonic
 - c** the 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 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 m s^{-1} .

- 10** An organ pipe is an air column closed at one end with an effective length of 75 cm. The speed of sound inside the pipe is 330 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?

You have learned what sound is, how it is produced in acoustic and natural circumstances, and what the limitations of our hearing are. That information is the basis upon which sound engineers develop recording and playback systems.

Components of a sound system

Any audio system will include a series of key components: inputs, an amplifier, tone controls and loudspeakers.

- **Input**—Most audio systems are designed to work with a number of different sources, including a microphone, radio tuner, compact disk and DVD player. A switch connects the required input to the amplifier. The recordings on a CD, cassette or any other medium, as well as most sounds you hear on the radio, have at some stage been recorded using a microphone.
- **The amplifier**—The centre of any sound system, the amplifier should increase the size of the input signal with as little distortion as possible. The output waveform should simply be a bigger version of the original. The gain should be the same for all frequencies, so a graph of gain against frequency should produce a straight line—a linear frequency response curve.
- **Tone controls**—Often an equaliser is incorporated in a sound system to enable the increase or suppression of particular frequencies. The adjustment should make the reproduced sound more like the original, free of the effects of frequency distortion from the amplifier or venue.
- **Loudspeakers**—Loudspeakers convert the electrical signals of the amplifier back into sound waves. A variety of acoustic technologies are used to provide a sound as true to the original as possible. Some examples are specific speakers for particular frequency ranges, tuned 'ports', enclosures and baffles.

While there is a lot involved in the whole sound system, the aim is simple: to reproduce the original sound as faithfully as possible. Every component of the audio system is equally important. A high-fidelity—or hi-fi—sound system gets its name from its ability to accurately reproduce sounds in a wide range of frequencies. Ideally, a high-fidelity system will have a flat linear response curve.

This section takes a look at the first and last links of the system.

The first link: Microphones

While a salesperson might try to persuade you otherwise, a high-fidelity sound system would be useless if you wanted to play sound recorded using the simple carbon microphone of a telephone.

A wide range of microphones has been developed, each with a specific purpose and capable of responding in a particular way to the sound waves around it. The role of a microphone is to faithfully produce a varying potential difference (or voltage) with relative amplitudes over time that are exactly proportional to the variations in the incident sound. In technical terms, a microphone is an electro-acoustic device containing a transducer, actuated by sound waves and delivering essentially equivalent electrical signals.

Physics file

The term 'transducer' is widely used for many fields and applications. In simple terms a transducer—such as a microphone—is a device that receives waves (whether they be electrical, acoustical or mechanical) from one source, and supplies related waves (not necessarily of the same type as the input) to one or more other systems. If the transducer gets its energy solely from the input waves, it is called 'passive'. If additional energy from a source other than the input wave is required to power the transducer, it is called an 'active transducer'.

Classes of microphones include pressure, pressure-gradient or velocity, combination pressure and velocity, and wave-interference microphones. A class of microphones may contain one of the following transducers: carbon, ceramic, condenser, moving coil, inductor, ribbon, magnetic, electronic and semiconductor. The functioning of each type depends largely on electrical and electromagnetic effects, which you have encountered when studying electric power.

The choice of microphone will depend on its application—simple voice recording, high-fidelity studio recording or directional recording of particular sounds without background noise. It will also be affected by price constraints, as high-fidelity microphones cost several hundred dollars. The vast range of microphone types available today means that there is a specialist design available for every task.

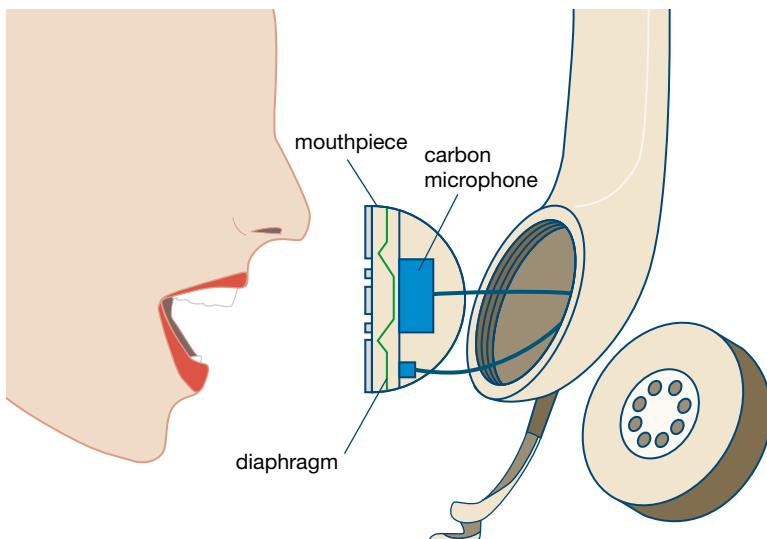


Figure 15.38 A telephone microphone only needs to respond to the relatively limited frequency range of the human voice. It can simply consist of carbon granules packed between two thin metal plates. Sound waves from your voice compress and decompress the granules, changing the resistance of the carbon and modulating the current flowing through the microphone. This type of microphone is referred to as a *pressure microphone*, as it depends on pressure variations in the air to operate. For improved performance in modern communication devices, pressure microphones have largely been replaced by dynamic, magnetic and electret-condenser microphones with built-in amplifiers.

Piezoelectric microphones

A piezoelectric microphone contains a transducer element that generates a voltage when it is mechanically deformed. ('Piezo' comes from the Greek word for pressure.) The induced voltage is proportional to the displacement. Originally a particular asymmetric salt crystal was used, hence the name 'crystal microphones'. Now piezoelectric microphones contain newer ceramic materials, which are more resistant to environmental extremes. They are also called 'ceramic microphones'.

The most common construction today is shown in Figure 15.39. The piezoelectric material is mounted as a cantilever and actuated by the drive pin. The diaphragm itself is made of aluminium or polyester film.

The advantage of the ceramic microphone is that its output voltage is sufficient to drive a high-impedance input amplifier directly. The frequency response is uniform through to values above 10 000 Hz and it is stable over time and over a wide range of environmental conditions. The cost is also low. For these reasons, piezoelectric microphones were widely used for home recording in the early days of valve amplifiers, until the advent of solid-state transistor-based equipment created a demand for other types of microphones that would work with modern amplifiers.

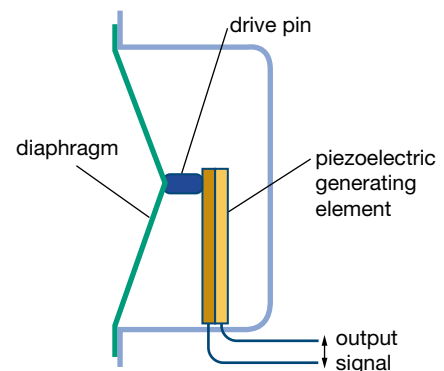


Figure 15.39 Typical construction of a piezoelectric microphone.



Figure 15.40 Dynamic or moving coil microphones remain a popular choice for vocalists and musicians alike.

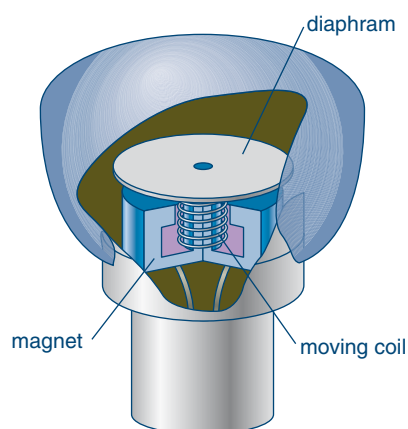


Figure 15.41 Typical construction of an electrodynamic microphone.

Electrodynamic microphones

The moving coil microphone has been one of the most popular microphones for high-quality recording for a long time. This type is capable of responding well to sounds in the whole range of human hearing, from around 20 Hz to 20 kHz. In 1831 Michael Faraday discovered that if the strength of a magnetic field near a wire is changing, an electromagnetic field (EMF) is induced and a voltage produced across the wire. If the wire is part of a complete circuit, there will be an induced current.

In a moving coil, or (electro)dynamic, microphone, a small diaphragm is attached to a coil of wire as shown in Figure 15.41. When someone speaks into the microphone, the diaphragm moves up and down. The coil will move between the poles of a permanent magnet inducing a voltage in the coil. This voltage will be equal to:

$$E = Blv$$

where E = the induced voltage (V)

B = the magnetic field strength (T)

l = the length of the conductor in the air gap (m)

v = the velocity of the coil (m s^{-1}).

If the velocity of the coil varies due to the frequency and amplitude of the sound waves incident on the diaphragm, then the induced voltage will also vary. The mechanical resonance of coil and diaphragm can be damped—electrically or mechanically smoothed—to give a uniform frequency response throughout the required range.

Velocity microphones

Pressure gradient, or velocity, microphones also work on the principle of induction. Rather than a coil, a velocity microphone has a thin metal ribbon suspended between the poles of a permanent magnet. The ribbon vibrates in response to the sound waves, and the EMF induced in the ribbon is proportional to the speed of vibration. Magnetic fine-mesh steel screens are placed on both sides of the ribbon to provide resistance damping of the ribbon (reducing vibrations from other sources) and to protect it from dirt.

The ribbon is tuned to resonate at around 30 Hz. Typically, the achievable frequency response for a velocity microphone is 30–15 000 Hz. The most important characteristic of this type of microphone is that it can easily be adapted to function as a directional microphone: thanks to the shielding screens, the ribbon can be exposed to sound waves from a particular direction only.

Electret-condenser microphones

A condenser microphone operates on the basis of variations in internal capacitance (the storage of electrical charge). As the diaphragm vibrates with the sound wave, its motion is translated into a voltage by the relative proximity of the charges, raising and lowering the capacitance.

The foil, shown in Figure 15.42, is selected on the basis of a compromise between electrical and mechanical properties. Polycarbonate films are one set of suitable materials. Typically, one side of the film is coated with a thin (50 nm) layer of a conductive metal, such as aluminium, gold or nickel. The foil is then heated and charged with a high DC potential. A well designed

electret 'capsule' is small in size and can retain its full charge and sensitivity for 10 years; it will take up to 100 years before its performance is really compromised.

On the negative side, the plastic foil cannot withstand the tensions that are needed for high resonant frequencies. Attempts to stiffen the foil can lead to short-term instability, limiting the high-frequency response and stability of the microphone. The great advantage, however, is that the foil can be made very cheaply in automated factories.

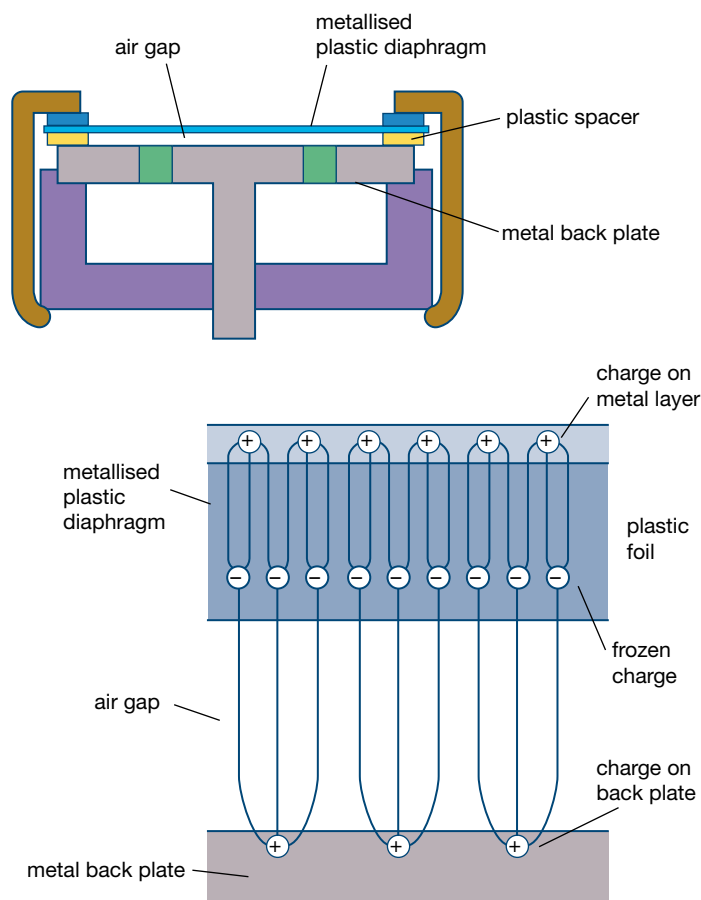


Figure 15.42 The simplest form of electret microphone is referred to as the 'charged diaphragm' type. The frozen charge and the charge on the metal back plate produce an electric field in the air gap between them.

Physics file

In any transmission of sound from a medium of one density to another medium of different density, some of the sound will be reflected, some absorbed and some transmitted. The proportion of a sound that is reflected, transmitted or absorbed depends on the relative *acoustic characteristics* of the media in question. If the media are similar, most of the sound will be transmitted. If the media are very different, most will be reflected. Even an open window can reflect a small proportion of sound, because of a difference in temperature inside and outside the room.

The last link: Loudspeakers

Although they may look very different, a moving coil microphone and a moving coil—or dynamic—loudspeaker are very similar. Each has a cone fixed to a coil of wire moving in a 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.



PRACTICAL ACTIVITY 51

Frequency response of a loudspeaker

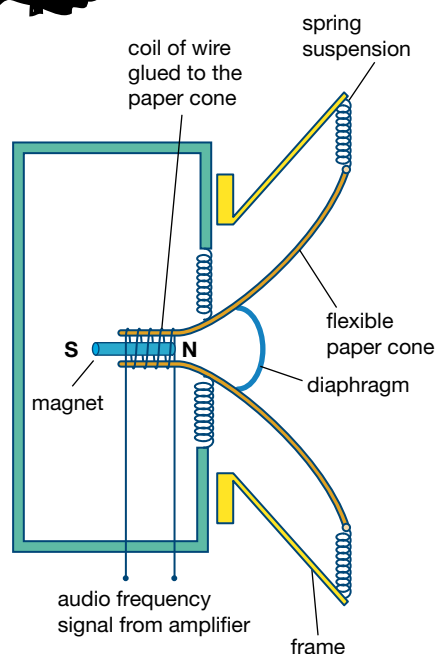


Figure 15.43 The principle on which a dynamic, or moving coil, loudspeaker operates. The changing current through the coil moves the magnet to and fro, making a paper or plastic cone vibrate, which in turn causes the air in front of it to vibrate. This can be clearly seen when the grill is removed from 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, as depicted in Figure 15.3.

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. If a current moving through a coil in one direction makes a magnet move to the left, reversing the current will cause it to move to the right. Figure 15.43 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 15.43 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 15.44 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.

Speaker damping

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.

Enclosures

When a speaker cone moves forward, the front surface sends out a compression wave. But at the same time, the rear of the cone is creating a rarefaction. At low frequencies (less than 200 Hz), diffraction effects cause the sound waves from the back of the loudspeaker to bend around the outer rim of the speaker and cancel out the sound waves from the front surface. To remedy this, the speaker is mounted in a box filled with some absorbent material. Sounds coming from the back surface of the cone are thus contained and absorbed, so they cannot interfere with the sound waves from the front surface.

Usually the various speakers making up the left channel of a stereo output are mounted together in one enclosure, and those making up the right in another, although the higher frequencies don't suffer as much from diffraction effects. Some modern systems are now keeping speakers separate to allow more 'tuning' of the listening environment. Either way, the supplied frequencies are filtered and each range is directed to the appropriate speaker: the high-frequency signals are sent to the tweeters and the low frequencies to the woofer or sub-woofer.

Baffles and ports

Designers have gone to considerable lengths to stop the unwanted sound from the back of the speaker superimposing with that from the front. Placing the loudspeaker in a large baffle will always improve the production of low frequencies, because of the increased distance from the back to the front of the speaker. The 'doof-doof' of car stereo systems heard way down the street is a good example. A large, or even infinite, distance—termed an *infinite baffle*—is desirable but hardly practical. Early solutions involved boxes or flat surfaces which shielded the front of the speaker. However, resonance also played its part, causing even more problems.

A successful early solution is shown in Figure 15.45. The closed box is modified by the inclusion of a carefully designed opening in the front. Through the 'vented' or 'ported' enclosure (also referred to as a bass-reflex monitor), sound from the back of the speaker can be added to that from the front without cancelling it. If the port is carefully designed, it acts like a second diaphragm driven by the back side of the speaker. It can add an octave or more to the system's low-end frequency response.

One key to the successful design of a port is to make sure the enclosure resonance matches that of the speaker itself. The process reverses the phase of the backwave, resulting in radiated sound that is in phase with the sound from the front of the speaker. Many other systems have been developed over the years, but ported enclosure remains a favourite.

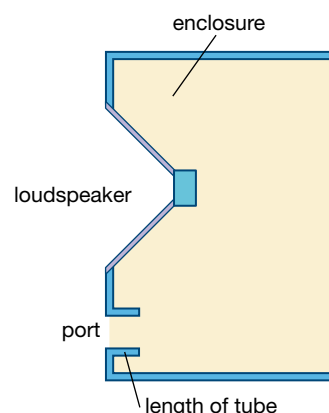


Figure 15.45 The simplest and most effective form of ported enclosure uses a hole for increased efficiency and extended low-frequency performance.

Physics file

The performance of even the best speaker system depends on the environment in which it is operated. The way the speaker interacts with its environment is complex, but one of the first considerations is the distance from each speaker to the wall, ceiling or floor. A speaker mounted in the centre of a room radiates low frequencies in all directions. Mounting the speaker near a wall allows the low frequencies to be reflected by the wall, which increases the low-frequency output by around 3 dB. This can have either positive or negative effects.



15.7 summary

Recording and reproducing sound: The first and last links

- Sound fidelity refers to the ability of a sound reproduction or recording system to accurately reproduce the original sound.
- Microphone and loudspeaker designs are based on simple electrical and electromagnetic effects.
- A piezoelectric microphone contains a transducer element that generates a voltage when it is mechanically deformed. The frequency response is uniform through to values above 10 000 Hz and it is stable over time and over a wide range of environmental conditions. The cost is also low. It is not as popular now as better options exist for modern amplifiers.
- A moving coil—or electrodynamic—microphone operates on the principles of induction. It has a diaphragm that causes a coil to move between the poles of a magnet, creating an alternating induced voltage. It is suited to high-quality recording.
- An electret-condenser microphone operates on the basis of varying the internal capacitance. It is not suitable for high resonant frequencies but is cheap to make.
- A velocity microphone operates on the principles of induction. A metal ribbon vibrates between the poles of a magnet. It can be easily adapted as a directional microphone.
- A vast range of other microphones have been developed for specific purposes.
- A moving coil—or dynamic—loudspeaker has a cone that is made to vibrate by the movements of a coil which undergoes an alternating current while sitting in a strong radial magnetic field.
- Specialist speakers have been developed for particular frequency ranges.
- Enclosures limit the diffraction effects that can occur for low frequencies, where sound waves coming round the back of the speaker can cancel out waves from the front.
- Baffles limit resonance effects inside the speaker enclosure. Tuned ports can direct this energy to the front, resulting in increased loudspeaker performance.



15.7 questions

Recording and reproducing sound: The first and last links

- 1 Describe the greatest difference between a moving coil microphone and a moving coil loudspeaker.
- 2 A loudspeaker can be described as a microphone working backwards—producing sound from electrical energy, rather than converting sound to electrical energy. This means that a loudspeaker could also be used as a microphone. What size moving coil loudspeaker would make the best microphone? Explain your answer.
- 3 Many loudspeaker enclosures include multiple speakers designed to accurately reproduce particular ranges of frequencies. Tweeters are used to reproduce high frequencies. Why would they be of no use for producing low frequencies?
- 4 A moving coil loudspeaker is attached to an oscilloscope for a simple experiment. It is then dropped onto its face from a short distance above a bench. Describe and explain what will be seen on the oscilloscope.
- 5 Explain the benefits of using a cylindrical magnet in the design of a moving coil loudspeaker.
- 6 Carbon microphones are widely used in home telephones, but they are deemed unsuitable for use in making high-fidelity recordings. Explain why.
- 7 A loudspeaker enclosure can include tweeter, mid-range and woofer speakers to more accurately reproduce particular frequencies. Draw a simple diagram of a frequency response curve to explain how the speakers should work together to produce a broad frequency range.
- 8 A sound studio wishes to make a high-quality recording of a new band. Based on the descriptions included in this study, which type of microphone should the studio choose? Why?
- 9 Describe the basic principle of the moving coil microphone.
- 10 Why is a microphone referred to as a ‘transducer’? Would it be reasonable to also refer to a loudspeaker as a transducer?



chapter review

In these questions, assume that the speed of sound in air is 340 m s^{-1} .

Multiple-choice questions

The following information applies to questions 1–6.

A microphone is placed in front of a sound source. The output is fed into a digitiser connected to a computer. X and Y are the traces from two different sounds shown on the screen of the computer. The sampling settings are identical for each trace.



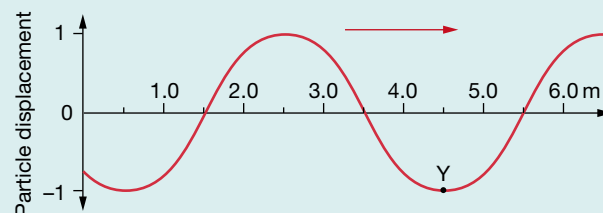
- Which of the following do the traces represent?
 - Pressure variation versus time of a transverse wave
 - Displacement of molecules versus time of a transverse wave
 - Pressure variation versus time for a longitudinal wave
 - Movement of individual molecules in the air up and down directly representing the sound wave
- What is the ratio of the amplitudes $A_X : A_Y$?
 - 1:1
 - 2:1
 - 1:2
 - 1:4
- What is the ratio of the frequencies $f_X : f_Y$?
 - 1:1
 - 2:1
 - 1:2
 - 1:4
- How many rarefactions of wave X are represented in trace X?
 - 1
 - 2
 - 3
 - 4
- How many compressions of wave Y are represented in trace Y?
 - 1
 - 2
 - 3
 - 4

- Which one (or more) of the following statements about the pressure variation in a sound wave as it relates to compressions and rarefactions is true?

- Pressure variation is at a maximum halfway between a compression and a rarefaction.
- Pressure variation is at a minimum halfway between a compression and a rarefaction.
- Pressure variation is zero at a compression and it is at a maximum at a rarefaction.
- Pressure variation is zero at a rarefaction and it is at a maximum at a compression.

The following information relates to questions 7–10.

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 340 m s^{-1} .



- What is the wavelength of the sound wave?
 - 1.0 m
 - 2.0 m
 - 4.0 m
 - 8.0 m
- Which arrow describes the direction of motion of a molecule at point Y just after this particular time?
 -
 - ←
 - ↓
 - ↑
- Which arrow below describes the direction of transfer of acoustic energy by this wave?
 -
 - ←
 - ↓
 - ↑
- Which of the following properties of sound is independent of the source producing the sound?
 - frequency
 - amplitude
 - speed

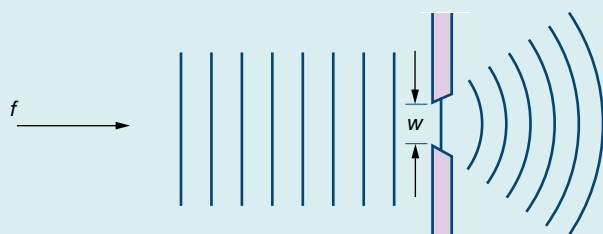
- 11 Sound waves of frequency f are being diffracted as they pass through a narrow slit of width w . The amount of diffraction can be increased [Choose one or more answers.]:

- A by increasing f
- B by increasing w
- C by decreasing f
- D by decreasing w .

Extended-answer questions

The following information relates to questions 12 and 13.

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.



- 12 What is the lowest frequency of sound that will produce resonance in the tube?
- 13 What frequency of sound will cause the tube to resonate at its third harmonic?
- The following information relates to questions 14–16.
- A loudspeaker is emitting a sound with an intensity, 10 m in front of the speaker, of $1.6 \times 10^{-6} \text{ W m}^{-2}$.
- 14 Calculate the acoustic power of the loudspeaker.
- 15 What would the sound intensity be 20 m in front of the speaker?
- 16 How much acoustic energy does the loudspeaker emit in 10 min?
- 17 Explain the difference between perceived loudness and intensity.
- 18 Two Occupational Health and Safety officers record the sound level in a particular factory at 63 dB in an area where the safety limit is 60 dB. The junior of the two officers argues that this is only about 5% above the limit and is not a significant breach of the safety code. Discuss the merits of this argument.
- 19 A large group of healthy young people have their hearing tested. They listen to recorded sounds of various frequencies, the sound level of which is kept at a constant 60 dB. The frequency range of the sounds is from 100 Hz to 10 kHz. What would be the most common effect on perceived loudness for the participants in this test?
- 20 The test from Question 19 is repeated with a signal generator attached to a small tweeter speaker. Would this remain a fair test? Explain your answer. What type of speaker should be used for the test to be fair?
- 21 Explain the purpose of baffling in a loudspeaker enclosure.
- 22 Which type of microphone would be most suitable in the manufacture of mobile phones? Explain your answer.

Chapter 1 Motion

1.1 Mechanics review

1 S, V, S, V, S, V, V, S 2 C

3 a (i) 40 m (ii) 160 m (iii) 160 m b 200 m north
c 360 m d 13 m s⁻¹ e 0 m s⁻² 4 33 s

5 a 9.80 m s⁻¹ down b 19.6 m c 14.0 m s⁻¹

d after 4.0 s e (i) 9.80 m s⁻² down (ii) 9.80 m s⁻² down

6 a 98 N b 69 N

7 a 1.3 m s⁻¹ b 1.1 m s⁻² c 2.7 m s⁻¹ d 1.9 m s⁻¹

8 a 6.26 m s⁻¹ b 6.26 m s⁻¹ c 5.42 m s⁻¹ d tomato;
-6.26 m s⁻¹ e golf ball; 11.7 m s⁻¹ up

9 a 16 m b 4.0 s c 2.0 m s⁻¹ down ramp

10 a 103 m s⁻¹ south 14.0° west b south 14.0° east

1.2 Newton's laws of motion

1 When the car is at rest, its tendency as described by Newton's first law is to remain at rest. A large force is needed to overcome its inertia and start the car rolling. Once the car is rolling, however, its tendency is to continue rolling. Only a small pushing force, to overcome resistance forces, is needed to maintain the car's motion.

2 No, Phil stayed where he was as the tram moved forward. This is an example of the law of inertia. Objects will remain at rest unless an unbalanced force acts to change the motion. 3 0.098 N upwards

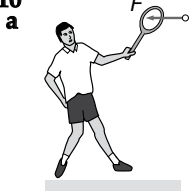
4 C 5 a 45 N b 165 N

6 a 1.5 m s⁻² b 120 N c 60 N

7 a 0 N b 66 N c 66 N 8 a equal b opposite

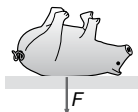
9 The trolley will move. The horizontal forces acting on the trolley (pushing force and friction) are unbalanced. The pushing force is greater than friction so the trolley will accelerate forwards.

10



Force exerted on racquet by ball

b



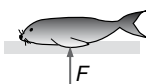
Force exerted on ground by pig

c



Force exerted on ground by wardrobe

d



Gravitational force of attraction that seal exerts on Earth

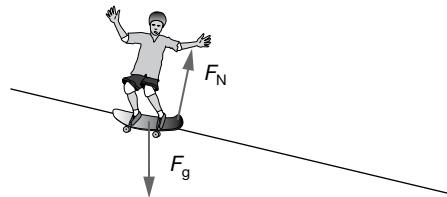
1.3 The normal force and inclined planes

1 A 2 C 3 490 N up the hill 4 4.9 m s⁻²

5 Acceleration is unchanged at 4.9 m s⁻²

6 a 250 N b 530 N down the incline c 8.9 m s⁻²

down the ramp 7 a 9.4 m s⁻¹ b F_g (vertical) and F_N (perpendicular to slope)



8 530 N up the ramp 9 a 0.47 N b 3.8 m s⁻¹

10 a $F_N = 0$ b $F_g < F_N$ c $F_g \ll F_N$

1.4 Projectile motion

1 a 1.0 s b 20 m c 9.8 m s⁻² down d 21 m s⁻¹
e 22 m s⁻¹

2 a 10 m s⁻¹ b 4.4 m s⁻¹ down c 11 m s⁻¹ at 24° to the horizontal d 0.45 s e 4.5 m f Diagram should show vertical gravitational force F_g acting.

3 a 24 m s⁻¹ b 24 m s⁻¹ c 24 m s⁻¹

4 a 14 m s⁻¹ up b 4.2 m s⁻¹ up c 5.6 m s⁻¹ down

5 a 1.43 s b 10 m c 9.8 m s⁻² down

6 a At the maximum height b 24 m s⁻¹ c 1.43 s

d Diagram should show vertical gravitational force F_g acting.

7 a 2.9 s b 28 m s⁻¹ at 30° to horizontal c 69 m 8 C

9 a 6.5 m b 11 m s⁻¹ up c 45° d 11.4 m s⁻¹ e 13 m at an angle of 30° to the horizontal f 2.3 s g 26 m

10 a 22.9 m s⁻¹ b 11.0 m c 3.57 s

Chapter review

1 a -0.5 m s⁻¹ b 6.5 m s⁻¹ up

2 a The force that the trampoline exerts on David is initially very small, but increases in strength until the point of maximum extension. The reverse process then occurs. This force is usually called the normal force.

b D 3 a 10.2 s b 510 m c 9.8 m s⁻² down

4 a 9.3 m s⁻² b 2.8 m s⁻¹ c 1.1 kN 5 a C b D

6 a 0.64 s for both b 8.0 m s⁻¹ c 9.8 m s⁻¹ d 1.6 m

7 62.3° 8 4.6 m s⁻¹ 9 D 10 8.4 m 11 a 4.0 m s⁻¹

b 6.9 m s⁻¹ c 0.71 s d 3.9 m e 1.6 s f 6.4 m

12 4.0 m s⁻¹ 13 16 J 14 9.8 m s⁻² down 15 C

16 A 17 2.7 J 18 C 19 2.1×10^{-3} N

20 9.2×10^3 . Clearly, the drag force is much smaller than the weight force.

Chapter 2 Collisions and circular motion

2.1 Momentum and impulse

1 The mosquito

2 Pavithra; she has momentum of 160 kg m s^{-1} compared to Michelle's 150 kg m s^{-1} .

3 a $2.7 \times 10^{-23} \text{ kg m s}^{-1}$ **b** $1.8 \times 10^{29} \text{ kg m s}^{-1}$
c 0.23 kg m s^{-1} **d** D

5 a This extends his stopping time and so reduces the size of the force that acts to stop him. Using $F\Delta t = \Delta p$, it can be seen that if the momentum change occurs over a long time interval, the force that is acting must be reduced in size. **b** His momentum would change over a very short time interval. It would require a very large force to do this, which would be painful (or damaging) for Vijay.

6 a -2.0 m s^{-1} **b** 18 m s^{-1} upwards **c** 1.4 kg m s^{-1} upwards **d** 1.4 N s upwards

7 a 29 N up **b** 30 N up **c** 30 N down

8 a 500 N forwards **b** 500 N backwards **c** 5.0 N s forwards **d** 5.0 kg m s^{-1} forwards **e** 50 m s^{-1}

9 The padding extends the time over which the players lose their momentum and are brought to a stop. The force that is acting to stop them must therefore be reduced in size.

10 a (i) equal for both cars **(ii)** equal **(iii)** B

b $F_A = 5.0 \times 10^3 \text{ N}$; $F_B = 2.0 \times 10^4 \text{ N}$ **c** Crumple zones reduce the magnitudes of the forces that act on the occupants of a car during a collision. This will result in fewer or less serious injuries. **d** This will further extend the time taken for a passenger to come to a stop. Using $F\Delta t = \Delta p$, this will result in smaller forces acting on the person as they come to a stop.

2.2 Conservation of momentum

1 a $1.0 \times 10^4 \text{ kg m s}^{-1}$ east **b** $1.0 \times 10^4 \text{ kg m s}^{-1}$ west **c** 0

2 a 0 **b** It hasn't gone anywhere. The vehicles had a total of zero momentum before the collision and so there is zero momentum after the collision.

c $4.0 \times 10^4 \text{ kg m s}^{-1}$ west **d** $4.0 \times 10^4 \text{ kg m s}^{-1}$ east

3 12 m s^{-1} right **4** 5.0 m s^{-1} **5** 22 m s^{-1} right

6 Mary is correct. As the water spills from the tanker, the water continues to move forwards at 5.0 m s^{-1} . Because the water keeps its horizontal momentum so too will the tanker retain its momentum. There is no transfer of momentum between the tanker and the water, so each will continue to travel forwards at 5.0 m s^{-1} .

7 2.5 m s^{-1} **8** C **9 a** B **b** A

10 a 190 kg m s^{-1} right **b** 3.8 m s^{-1} right
c 3.8 m s^{-1} right

11 The footballer's momentum has reduced to zero. His momentum has been transferred to the Earth, making it move very slightly.

12 No, the diver is not an isolated system.

An unbalanced gravitational force is acting on her.

2.3 Work, energy and power

1 a $7.1 \times 10^5 \text{ J}$ **b** $3.9 \times 10^5 \text{ J}$ **c** 16 kW

2 a 38 N m **b** 25 N m **c** 13 N m

3 a 13 J **b** 25 J **c** friction

4 Zero. The vertical component has had no effect on the energy of the dog. **5** 1.4 s

6 a 8.1 J **b** 8.1 J **c** This collision is elastic since there is no loss of kinetic energy during the collision. **d** No. Truly elastic collisions only occur at the atomic level.

7 a 0.050 N m **b** During compression, the girl does work to alter the shape of the ball. During release, the stored energy is released in restoring the ball to its original shape. **c** 0.10 W **8** 4.0 cm

9 a $3.9 \times 10^3 \text{ J}$ **b** $2.0 \times 10^3 \text{ J}$ **c** This is due to energy losses caused by friction. **d** $3.9 \times 10^3 \text{ N m}$ **e** $1.9 \times 10^3 \text{ J}$ **10 a** Yes **b** inelastic **c** 2.0 m s^{-1} in opposite directions

2.4 Hooke's law and elastic potential energy

1 C **2 a** 200 N **b** $1.0 \times 10^4 \text{ N m}^{-1}$ **c** 2.0 N m **d** 0.50 J

e The graph should show that $U_s \propto x^2$.

3 a A: $2.0 \times 10^4 \text{ N m}^{-1}$; B: $1.0 \times 10^4 \text{ N m}^{-1}$;
C: $5.0 \times 10^3 \text{ N m}^{-1}$ **b** C, B, A **c** A: 1.0 J ; B: 0.50 J ; C: 0.25 J

4 0.71 **5 a** 15 N m^{-1} **b** $1.8 \times 10^{-2} \text{ J}$

6 a 0.16 J **b** 0.23 kg **7 a** 2.8 J **b** 2.8 J

8 a 0.12 m **b** 2.5 J

9 a 7.2 J using 14.5 squares under the graph **b** It is being stored as elastic potential energy in the bow.

10 a 7.2 J **b** There is no transformation of energy due to friction as the string returns to its unextended position.

2.5 Circular motion

1 a A, D **b** Her natural tendency has been to travel forwards at a tangent owing to inertia. The car has turned to the left, giving the illusion that she has been thrown to the right.

2 a 8.0 m s^{-1} **b** 8.0 m s^{-1} south **c** 7.0 m s^{-2} east

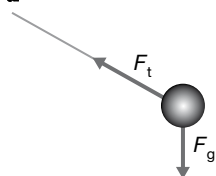
3 a $8.4 \times 10^3 \text{ N}$ east **b** Friction between the tyres and the road **4 a** 8.0 m s^{-1} north **b** west

5 The force needed to give the car a larger centripetal acceleration will eventually exceed the maximum frictional force that could act between the tyres and the road surface. At this time, the car would skid out of its circular path.

- 6 **a** 2.7 m s^{-2} towards the centre of the circular path
b Unbalanced; this is why she has an acceleration.
c 130 N **d** Friction between the skate blades and the ice
7 a 0.50 s **b** 10 m s^{-1} **c** 130 m s^{-2} towards centre of circle **d** 320 N **e** The tension in the wire
f Parabolic in a vertical plane tangential to the circle
8 28 s **9** 5.0 N
10 The sideways force that the air exerts on the plane

2.6 Aspects of circular motion

- 1 a** west **b** south-west **c** north
2 a south **b** south-east **c** west
3 a south **b** north-east **c** north-west
4 east **5** C
6 a 3.70 m s^{-1} **b** 17.1 m s^{-2} radially inwards **c** 0.43 N
7 a



- b** 0.490 N
8 a 1.2 m **b** Gravity, which is vertical, and tension in the rope towards point A **c** towards B **d** 170 N
e 2.6 m s^{-1} **9 a** 4.9 kN **b** 22°
10 The greater speed would make the car travel higher up the track. The driver would have to turn the wheels slightly towards the bottom of the track so as to create a sideways force of friction towards the bottom of the track.

2.7 Circular motion in a vertical plane

- 1 a** Acceleration is equal in magnitude at all times.
b At bottom of path **c** At top of path
d At the bottom of its circular path
2 3.8 m s^{-1}
3 a Normal force from road, weight
b $1.3 \times 10^3 \text{ N}$ down **c** Yes. When the driver is moving over a hump, the normal force is less than her weight mg . Her apparent weight is given by the normal force that is acting and so the driver feels lighter at this point.
d 36 km h^{-1}
4 a 31.4 m s^{-1} **b** 19.9 m s^{-1} **c** $8.3 \times 10^3 \text{ N}$ down
5 12.1 m s^{-1} **6** 200 N down **7** 31 m s^{-1} **8** 190 m s^{-1}
9 a 39 N **b** 120 N
10 When the ball is moving through X. It has an upwards acceleration and so tension is greater than the weight force. This larger tension in the wire means it is more likely to break.

Chapter review

- 1 a** 2.8 kg m s^{-1} down **b** C **c** D **d** $4.8 \times 10^3 \text{ N}$ up
2 a 225 J **b** 1.50 kJ **c** $1.28 \times 10^3 \text{ J}$
3 a 150 W **b** 128 W **c** 22.5 W
4 The section between 2.0 and 5.0 m. It is during this interval that the puck loses kinetic energy due to friction. **5** 1.0 N **6** A
7 a 0 **b** 200 kg m s^{-1} east **c** 200 kg m s^{-1} west
8 1.0 m s^{-1} west **9** 1.7 m s^{-1}
10 a 136 kg m s^{-1} west **b** 136 kg m s^{-1} east **11** D
12 1.0 m s^{-1} east **13** C **14** B **15** 8.4 J **16** A, C, D
17 a A **b** D **c** C **18** 13 s
19 a 5.0 m s^{-2} west **b** $7.5 \times 10^3 \text{ N}$ south
c $7.5 \times 10^3 \text{ N}$ east
20 C **21** 48 km h^{-1}
22 640 N. This is greater than the normal force that exists on a flat track (540 N).
23 a (i) 370 N up (ii) 620 N up **b** D

Chapter 3 Gravity and satellites

3.1 Newton's law of universal gravitation

- 1 a** $5.3 \times 10^{-12} \text{ N}$ **b** $6.7 \times 10^{-7} \text{ N}$ **c** $1.6 \times 10^5 \text{ N}$
d $2.0 \times 10^{20} \text{ N}$ **e** $5.9 \times 10^2 \text{ N}$ **f** $3.61 \times 10^{-47} \text{ N}$
2 a The term *universal constant* refers to the fact that the value of G (i.e. $6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$) has exactly the same value everywhere in the Universe. This means that Newton's law of gravitation is valid throughout the Universe. **b** Another very famous universal constant is the speed of light.
3 a Location is two Moon radii from centre. **b** 10 N
4 38 **5 a** $2.48 \times 10^4 \text{ N}$ **b** $2.48 \times 10^4 \text{ N}$ **c** 24.8 m s^{-2}
d $1.30 \times 10^{-23} \text{ m s}^{-2}$
6 10 Earth radii **7 a** 100 **b** 0.1R **8** distance = 0.91R

3.2 Gravitational fields

- 1** Mercury $g = 3.70 \text{ N kg}^{-1}$, Saturn $g = 10.4 \text{ N kg}^{-1}$, Jupiter $g = 24.8 \text{ N kg}^{-1}$
2 a 300 N **b** 830 N **c** $2.0 \times 10^3 \text{ N}$
3 The density of the Earth is much greater than that of Saturn.
4 0.61 N kg^{-1} **5** 0.61 m s^{-2} **6** 0.25
7 $2.0 \times 10^{12} \text{ N kg}^{-1}$
8 $8.0 \times 10^6 \text{ N kg}^{-1}$
9 $8.0 \times 10^6 \text{ m s}^{-2}$
10 The gravimeter would respond to differences in the gravitational field strength in a particular region due to either large deposits of low density material (e.g. oil) or large deposits of high-density material (e.g. iron ore).
11 a 270 N **b** 0.54 m s^{-2} **c** 27 N **d** 0.54 m s^{-2}
e 0.54 N kg^{-1} **12** $3.4 \times 10^8 \text{ m}$

3.3 Satellites in orbit

1 D **2** Since the gravitational force of attraction between the Earth and the satellite is always perpendicular to the satellite's velocity, the speed of the satellite remains constant, and therefore its change in kinetic energy is zero.

3 510 N **4** The centripetal force is the gravitational force of attraction between the Earth and the satellite.

5 a $5.40 \times 10^{-2} \text{ m s}^{-2}$ **b** $4.38 \times 10^3 \text{ m s}^{-1}$ **c** 5.90 days

6 a 5.55 km s^{-1} **b** $2.52 \times 10^{-2} \text{ m s}^{-2}$ **7** $5.64 \times 10^{26} \text{ kg}$

8 a $1.54 \times 10^9 \text{ m}$ **b** 460 m s^{-1} **c** $1.38 \times 10^{-4} \text{ m s}^{-2}$

9 a (i) 1.66 **(ii)** 7.58 **b** 15.6 days

10 $7.0 \times 10^6 \text{ m}$ **11** The centre of the circular orbit of a satellite must be the centre of the Earth.

12 a 0.31 N kg^{-1} **b** 340 m s^{-1}

3.4 Energy changes in gravitational fields

1 C **2** The gravitational field strength increases.

3 The acceleration of the meteor will increase as given by the gravitational field strength. **4** A, B, C

5 a 9.2 N **b** $2.6 \times 10^6 \text{ m}$

6 a $8.0 \times 10^6 \text{ J}$ **b** $2.0 \times 10^7 \text{ J}$ **c** $2.8 \times 10^7 \text{ J}$ **d** 7.5 km s^{-1}

7 a 7.6 km s^{-1} **b** $5.7 \times 10^{11} \text{ J}$ **c** $5.8 \times 10^3 \text{ s}$

8 a 6.7 km s^{-1} **b** $4.4 \times 10^{11} \text{ J}$ **c** $8.5 \times 10^3 \text{ s}$

9 $\sim 2.6 \times 10^{11} \text{ J}$ **10** $1.7 \times 10^9 \text{ J}$

3.5 Apparent weight and weightlessness

1 a 1.5 N **b** 0.15 kg **2** zero

3 The motion is unchanged.

4 a 490 N **b** 0 N **c** 0 N

5 a The weight of the girl is not zero, but she perceives her weight as zero because she feels zero reaction force from the seat during the plummet. True weightlessness only occurs when the gravitational field strength is zero. This only happens in deep space.

b It must be equal to g .

6 a 880 N **b** 3.2 kN **c** 4.0 kN

7 a 8.7 N kg^{-1} **b** 520 N

8 a 7.7 km s^{-1} **b** 8.7 m s^{-2} towards the centre of the Earth. **c** 520 N towards the centre of the Earth.

9 This comment is incorrect since the gravitational field strength is not zero. A better comment would be 'The astronaut is experiencing apparent weightlessness'.

10 The spacecraft and the astronaut are both moving with an acceleration that is equal to the gravitational field strength and so the astronaut will experience zero normal force from the spacecraft.

Chapter review

1 $3.78 \times 10^8 \text{ m}$ **2 D** **3 a D b B c C d A e D**

4 16 **5 a** 11.1 N kg^{-1} **b C** **6** 27.4 days

7 a $3.39 \times 10^3 \text{ m s}^{-1}$

b $1.05 \times 10^{-3} \text{ m s}^{-2}$ towards Jupiter **c** 235 days

8 a C b This satellite will always be in the same place in the sky and can therefore transmit radio signals to any point that can see it. **c** $4.2 \times 10^7 \text{ m}$

9 a $2.43 \times 10^8 \text{ m}$ **b** 301 m s^{-1} **c** $3.73 \times 10^{-4} \text{ m s}^{-2}$ towards Mercury

10 a $2.8 \times 10^7 \text{ J}$ **b** $3.8 \times 10^7 \text{ J}$ **c** $1.9 \times 10^3 \text{ m s}^{-1}$

d 3.5 N kg^{-1} **11 a C b B c D** **12** $2.0 \times 10^{30} \text{ kg}$

13 9.0 N kg^{-1} **14 D** **15 C** **16** $3.5 \times 10^9 \text{ J}$

17 No; air resistance would have a major effect.

18 D **19** 25 days **20** $1.3 \times 10^{22} \text{ kg}$

Exam-style questions (Motion in one and two dimensions)

1 $4.3 \times 10^{10} \text{ J}$ **2** $3.3 \times 10^3 \text{ m s}^{-1}$ **3 a** $4.0 \times 10^4 \text{ N}$

b $8.1 \times 10^4 \text{ N}$ **4** It increases from 8.1 m s^{-2} to 9.2 m s^{-2} .

5 3.4 km **6 a A b** Section with largest acceleration

7 10–40 s, 60–80 s **8** never **9** unbalanced, balanced

10 12.6 m s^{-1} **11** 31.6 m s^{-2} **12** $1.89 \times 10^3 \text{ N}$

13 0.600 Hz **14** $5.0 \times 10^7 \text{ m}$ **15** 11.7 km s^{-1}

16 $1.16 \times 10^8 \text{ m}$ **17** 4.9 m **18** 9.8 m s^{-2} down

19 11 m s^{-1} **20 D** **21** 10 kg m s^{-1} south

22 10 N s south **23 a** $1.0 \times 10^3 \text{ N}$ south **b** equal

24 C, E **25** 60 N **26** $\sim 3.6 \times 10^7 \text{ J}$

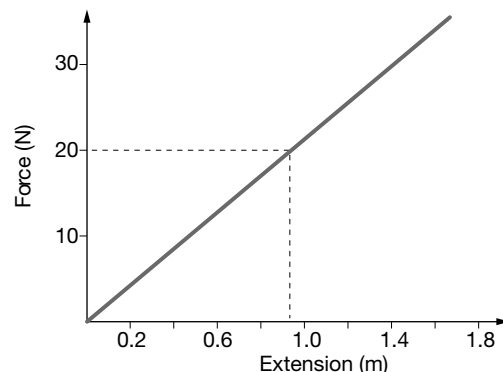
27 Have to find location where area under graph (after multiplying by mass) is equal to 40 MJ.

28 $2.0 \times 10^3 \text{ N m}$ **29** $1.4 \times 10^3 \text{ N m}$ **30** $1.4 \times 10^3 \text{ J}$

31 3.7 m s^{-1} **32** 600 J **33** 200 W **34** 0.42 J **35** 2.5

36 0.064 **37** 39.1

38



39 $\sim 20 \text{ N m}^{-1}$ **40** $2.3 \times 10^3 \text{ J}$ **41 C** **42** 8.7 m s^{-1}

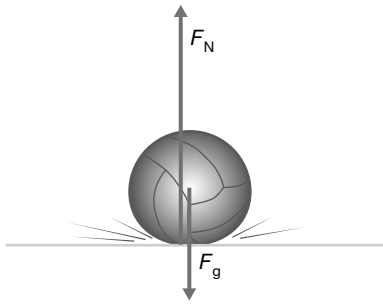
43 4.4 m s^{-1} **44** 200 kg m s^{-1} **45** 200 kg m s^{-1}

46 a inelastic, momentum, kinetic energy

47 a 18 m s^{-2} up **b** $1.5 \times 10^3 \text{ N}$

48 Apparent weight is greater than usual.

49



50 a C b Forces must act on different objects and be equal in size.

Chapter 4

4.1 Analysing electronic circuits

- 1 a** Lamp A gets brighter. **b** Lamp C turns off.
c Current increases. **d** Potential difference across lamp B increases. **e** Potential difference across lamp C decreases. **f** Total power increases.
2 a R_1 , R_4 and R_5 **b** R_2 and R_3 **c** R_1 , R_4 and R_5
3 a C b C
4 a Through $5\text{ k}\Omega$ resistor $I = 1\text{ mA}$ in direction P; through $2\text{ k}\Omega$ resistor $I = 1.5\text{ mA}$ in direction P.
b Through $5\text{ k}\Omega$ resistor $I = 1.14\text{ mA}$ in direction P; through $2\text{ k}\Omega$ resistor $I = 1.14\text{ mA}$ in direction P.

5

$R_1\text{ (}\Omega\text{)}$	$R_2\text{ (}\Omega\text{)}$	$V_{\text{out}}\text{ (V)}$
1000	1000	10
3000	1000	5.0
400	100	4.0
900	100	2.0
2.0	3.0	12

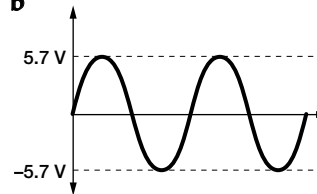
6

$R_1\text{ (}\Omega\text{)}$	$R_2\text{ (}\Omega\text{)}$	$R_3\text{ (}\Omega\text{)}$	S	$V_{\text{out}}\text{ (V)}$
200	200	100	open	10
300	100	25	open	5.0
50	50	75	open	15
100	100	1000	open	19
200	100	100	closed	0

- 7 a** $1600\text{ }\Omega$ **8** 0.5 W is $288\text{ }\Omega$; 1 W is $144\text{ }\Omega$ **9 a** same
b 1 W globe **c** 1 W globe
10 a 0.5 W globe **b** same **c** 0.5 W globe
11 a D b F c D d A **12 a** $48.8\text{ k}\Omega$ **b** 8.2 V
c 0.145 mA **d** 1.08 V across the 10 and $30\text{ k}\Omega$ resistors.

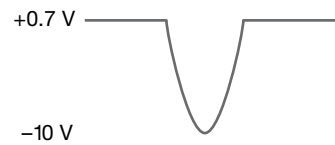
4.2 Diodes

- 1 a** 0 A **b** 1.3 mA **2 a (i)** Y and Z **(ii)** X **b** $16\text{ k}\Omega$
c Current through X = 3.2 mA ; current through Y = 4.0 mA **3 a A b B** **4 a** circuit 2 **b** circuit 1
5 a (i) B, C **(ii)** A
b

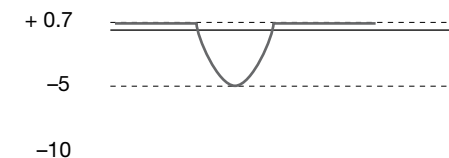


- 6 a (i)** Forward bias voltage that gives a normal operating current through the diode (i.e. where R_{diode} is very small). **(ii)** 0.3 V **(iii)** germanium **(iv)** $V_Y > V_X$
(v) From X to Y **b** Circuit 1, 13.5 mW ; circuit 2, 0 W .

7 a



b



- 8** $2\text{ k}\Omega$ **9** The diode is in reverse bias. Graph shows that almost no current flows when a reverse voltage is applied.
10 a The polarity of the battery has been reversed.
b 0.6 V **c** $1.2 \times 10^3\text{ }\Omega$

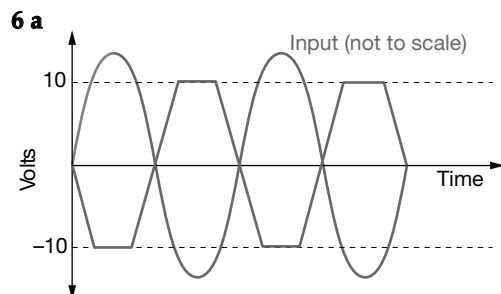
4.3 Amplification

- 1 a** Output voltage directly proportional to input voltage. **b** Ratio of change in output voltage to take the change in input voltage.
c Signal distortion when gain is non-linear. **d** Ratio of output current over input current. **e** Output voltage is 180° out of phase (i.e. opposite sign) with respect to input voltage.
2 a 8 V **b** If the amplifier was operating within its limits, an increase in the size of the variation of the input voltage Δv_{in} should produce a proportional increase in Δv_{out} . However, since the transistor has reached the limit of its input range, the output voltage has reached its minimum and maximum possible values and will not go any lower or higher. The output signal will be clipped.
3 Biasing transistor in the middle of its linear range

4 a For V_{in} below -0.5 V or above 0.5 V, amplifier is in non-linear region of operation. **b** 0.50 V **c** -20

5

Range of input voltage	Range of output voltage
100 to 200 mV	-2.0 to -4.0 V
0.25 to 0.50 V	-5.0 to -10 V
1.0 to 1.50 V	-10 to -10 V
20 to -20 mV	-0.40 to 0.40 V
0.80 to -0.80 V	-10 to 10 V



b No **c** The output waveform has been clipped and inverted. **d** The sound output would contain appreciable distortion. **7 D 8 A 9 B 10 D**

Chapter review

1 a R_4 **b** R_3, R_2 **c** R_4 **d** \mathcal{E}_2

2 a All four resistors in series. **b** Two resistors in parallel, and this combination in series with the other two resistors. **c** Four resistors in parallel. **d** Three resistors in parallel, and this combination in series with the other resistor.

3 $V = I(R + 3R)$ (i.e. $R_A + R_B$) $V_A = I(R)$ (i.e. R_A)
 $V_A/V = R/4R = 0.25$ **4 a** $R_A = 15 \Omega$; $R_B = 45 \Omega$ **b** 1.8 W

5 20°C **6 a** 75 V **b** 37.5 W

7

R_1 [Ω]	R_2 [Ω]	Switch	V_{out} [V]
1000	2000	open	60
2000	4000	open	60
4000	2000	open	50
8000	5000	closed	0

8 a At saturation, clipping occurs. The amplifier can no longer multiply the size of the signal by a consistent gain factor. **b** The output of the amplifier should be set to the middle of its possible output range so that optimal input-signal variation can occur without clipping.

9 a 0.23 A **b** 5.7 V

10 a Y **b** 1.9 k Ω **c** 40 mA **d** 98 V **e** 3.9 W

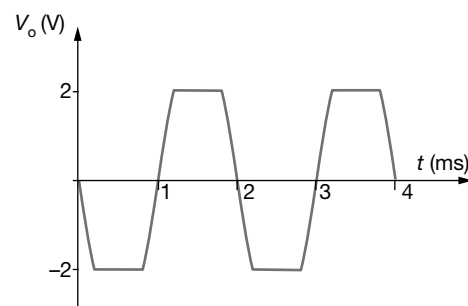
11 a $V_d = 0.7$ V, $I_d = 18.9$ mA **b** $V_d = -6$ V, $I_d = -100$ nA

12 a (i) 23 mA (ii) 1.4 mA (iii) 21.6 mA **b** (i) 5 mA

(ii) 5 mA (iii) 1 nA **13 a** 1.1 V **b** 2.1 V

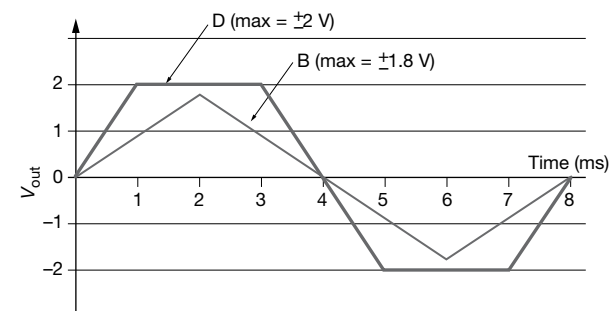
14 20 mV **15** Graph is inverted and varies between ± 1 V with the same period as input.

16

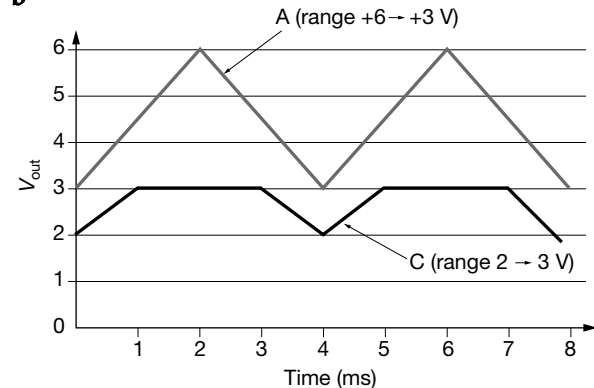


17 a A linear non-inverting amplifier **b** 10 **c** 4.0 V

18 a



b



19 a The sinusoidal signal is not distorted/clipped; therefore, the amplifier is producing linear gain. **b** 100

20 a 500 Hz **b** 500 Hz

Chapter 5

5.1 Photonics in telecommunications

1 B 2 B 3 A 4 D 5 D 6 A 7 D 8 C 9 A

5.2 Optical transducers

1 A 2 C

3 Advantage: higher optical gain; disadvantage: generally slower response time.

4 a 2 M Ω b 0.03 V c 3.4 k Ω

5 a Above minimum acceptable level

b 500–550 mW m⁻²

6 a 3.5 V b photoconduction 7 a 20 μ A b 4 W m⁻²

8 350 Ω 9 a (i) 117 Ω (ii) 150 Ω b same c Circuit ii because it draws less current from the battery.

10 a 500 Ω b 625 Ω c 10°C

5.3 Audio transmission via a light beam

1 A 2 B 3 D 4 B 5 C 6 D 7 A 8 A 9 B 10 C

Chapter review

1 B 2 D 3 B 4 a 0.2 W b 1% 5 8 W m⁻²

6 130 mW m⁻² 7 4 W m⁻²

8 a V_1 decreases, V_2 increases b ~ 0.64 V c 43 k Ω

d No 9 a 86 k Ω b V_0 increases to 10 V c $V_0 \approx 0$ V

Exam-style questions (Electronics and photonics)

1 a 2.0 A b The reading on the voltmeter is equal to the terminal voltage of the battery, i.e. 216 V. c 216 V

d 440 W e 432 W 2 220 V 3 a 98 Ω b 1.0 A

c (i) 80 V (ii) 10 V (iii) 8.0 V d 98 V e 100 W

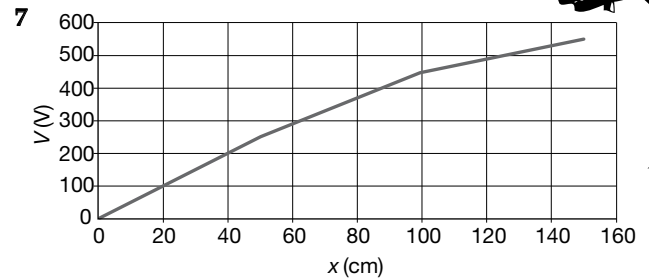
4 a 60 Ω b 2.0 A c $I_1 = 1.20$ A, $I_2 = 0.60$ A, $I_3 = 0.20$ A

d 240 W e 240 W

5

R_1 [Ω]	R_2 [Ω]	V_{out} [V]
1000	1000	10
3000	1000	5.0
400	100	4.0
900	100	2.0
2.0	3.0	12.0
400	100	4.0

6 a 0.5 b 2 c 4.0×10^{-18} J d 8.0×10^{-18} J e 100 V



8 a A material that has a greater resistance than metals, but less than insulators, e.g. Si and Ge. It conducts as a result of electrons gaining sufficient energy to break out of the crystal lattice. This may result from an energy input from light or heat. b n-type semiconductors are doped with an element such as phosphorus, which has 5 electrons in its outer shell (silicon has 4) and thus the extra one is mobile and becomes a charge carrier. p-type semiconductors are doped with an element such as aluminium which has 3 electrons in its outer shell (silicon has 4) and thus the missing electron acts as a 'hole' into which mobile electrons can jump, thus resulting in the hole moving and becoming a charge carrier. c The interface between a p-type and n-type semiconductor, where the extra electrons and holes will tend to recombine, forming a 'depletion layer', which does not conduct unless forward biased.

9 If the p-type material in a pn junction is made positive relative to the n-type material, holes will flow towards the junction from the p-type and electrons from the n-type; thus, they recombine and a current flows. Thus the pn junction is said to be forward biased. If the p-type is negative and n-type positive, i.e. the junction is reverse biased, the holes and electrons withdraw further from the junction, thus making it effectively an insulator.

10 a 20 mA b 20 mW c 200 d 0.10 W e 20 mA

11 a 1.5 V b 0.150 W c 0 d 3.5 V e 0 W f 35

g 0.35 W h 0.50 W 12 a 1.0 V b 20 mW c 0

13 a 22 mA b 11 mA c 11 mA

14 a 309.1 b 0.15 W c 220 mW 15 150

16 A diode will only conduct electricity when it is connected in forward bias. When a diode is connected in reverse bias its resistance is so great as to produce only a negligible current.

17 The voltage across the terminals of a diode beyond which the diode conducts strongly. For silicon $V_s \approx 0.7$ V, for germanium $V_s \approx 0.3$ V. 18 1.0×10^{-3} C

19 When the capacitor is fully charged, no more current flows in the circuit so no voltage drop across the resistor and all of the voltage across the battery appears across the capacitor.

$$V_c = Q/C = (50 \times 10^{-6}) / (100 \times 10^{-9}) = 500 \text{ V}$$

20 The amplified output voltage has the opposite sign to the input voltage (over the linear amplification range). For example, $V_{\text{in}} = +0.5 \text{ V}$, $V_{\text{out}} = -10 \text{ V}$ and for $V_{\text{in}} = -0.5 \text{ V}$, $V_{\text{out}} = +10 \text{ V}$.

21 a 0.50 V **b** -20 V

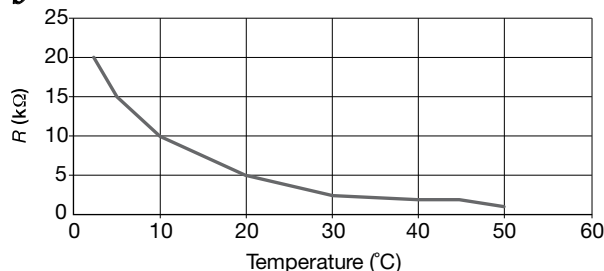
22

Peak input voltage	Output voltage range
100 mV	-2 to +2 V
0.25 V	-5 to +5 V
1.0 V	-20 to +20 V
20 mV	-400 to +400 mV
0.80 V	-16 to +16 V

23 a This is a non-inverting amplifier. This is because a positive voltage becomes a positive output voltage, and a negative input voltage becomes a negative output voltage. **b** 53 **c** 5.3 V

24 a A temperature-sensitive resistor whose resistance decreases as its temperature increases. It is usually constructed from Ge, Si or a mixture of various oxides. The resistance of a typical thermistor can range from $10 \text{ k}\Omega$ at 0°C to 100Ω at 100°C .

b



25 It isolates (or blocks) any DC component of the collector voltage of the first stage from the input to the second stage (but passes the AC signal voltage). This means that the DC biasing conditions of the second stage are not disturbed by the DC voltage of the first stage collector.

26 a 4 V **b** above **c** 0 V

27 a 200Ω **b** 87Ω **c** 79 mA **d** 2.6 V

28 a Like a normal diode but the pn junction is exposed to light which can release electron-hole pairs when the diode is reverse biased, thus allowing a small reverse current, which is proportional to the amount of light, to flow. **b** The current that flows when reverse biased while no light falls on the diode. It is very small, well under a microamp. **c** The current through a photodiode is only zero when there is a small positive voltage across it to oppose the photocurrent.

29 a $500 \text{ k}\Omega$ **b** $32 \mu\text{W}$ **c** $150 \mu\text{W}$ **d** $50 \mu\text{W}$
e $200 \mu\text{W}$

30 A transistor that has a photosensitive collector-base pn junction. The current produced by the photoelectric effect due to incident light is the base current of this device. The result is that an increase in the incident light intensity will produce an increase in collector current.

31 a $9.0 \mu\text{W}$ **b** $3.71 \times 10^{-19} \text{ J}$ **c** 2.42×10^{13} **d** $3.88 \mu\text{A}$
e 51.5

32 a A diode that will emit visible light when it is energised. When forward-biased, the recombination of holes and electrons takes place. During this recombination, some of the excess energy possessed by the unbound free electrons in the pn junction is converted into light energy. **b** 1.6 V **c** $1.6 \text{ k}\Omega$

d 22 mW **e** 100 mW

33 E

Chapter 6

6.1 Two principles that Einstein did not want to give up

1 Galileo realised that a force changed motion, not caused it.

2 Zero was not a special velocity; it was only relative to one frame of reference.

3 a 465 m s^{-1} **b** $30\,000 \text{ m s}^{-1} \pm 465 \text{ m s}^{-1}$ **c** The person at the Equator has acceleration towards the centre of the Earth (0.03 m s^{-2}). Both also have a very small acceleration (0.006 m s^{-2}) towards the Sun.

4 $2.3 \times 10^8 \text{ m s}^{-1}$, within 25%.

5 No, the escalator is an inertial frame of reference.

6 If pouring drinks in an accelerated frame of reference, the drink will not pour straight down.

7 The GPS unit picks up signals from satellites in the Earth's frame of reference.

8 a 370 m s^{-1} **b** 300 m s^{-1} **c** 340 m s^{-1} **d** 370 m s^{-1}

9 a higher **b** lower **c** higher **d** higher

10 To carry light waves just as air carries sound.

11 No; it was Galileo who introduced the principle of relativity. **12** At 0 m s^{-1} they both take 400 s, at 1 m s^{-1} Anna takes 417 s and Ben takes 408 s, at 4 m s^{-1} Anna takes 1111 s and Ben takes 667 s.

6.2 Einstein's crazy idea

1 Light stopped by a black surface disappears and turns into heat energy.

2 The changing electromagnetic fields would have to 'freeze', something that can't happen.

3 A, D **4 a** No, but it has negligible acceleration for most purposes. **b** At the poles.

5 a One example: if we held a pendulum it would

swing to one side. **b** The motion is not constant velocity.

6 A weight hanging on a spring balance would change force or direction if his ship began to accelerate.

7 a 15 m s^{-1} backwards **b** 3 m backwards **c** 0.2 s

8 a 0.1 s **b** 50 m s^{-1} **c** 1 m **d** 50 m s^{-1} **e** After a little less than 0.1 s (approx. 0.08 s) **9** A only

6.3 Time is not what it seems

1 2 s **2 a** 40 m s^{-1} **b** 0.75 s **c** 20 m s^{-1} , 1.5 s

3 a (i) 1.5 s for the backward flash and 0.75 s for the forward flash **(ii)** 2.25 s **b** 0.25 s longer

4 The flashes hitting the walls are simultaneous for Anna and Ben, but not for Chloe. The flashes returning to Anna and Ben are seen as simultaneous by all observers.

5 Chloe would agree with all Anna and Ben's measurements because she would add the velocity of the train to that of the ball or sound.

6 Chloe saw the light travel at 30 m s^{-1} in her frame. Because the train was travelling in the opposite direction, she saw the light meet the back wall in a time less than that given by $30 \text{ m} / 30 \text{ m s}^{-1} = 1 \text{ s}$. In Newtonian mechanics this would have meant that Anna and Ben saw the light travel at 40 m s^{-1} , but, as we have seen, that was not the case. No one saw light travel at any speed other than 30 m s^{-1} .

7 We assumed that lengths appeared the same to Chloe as to Anna and Ben.

8 a 1 m **b** $3.3 \times 10^{-9} \text{ s}$ **c** ct_c **d** $7.6 \times 10^{-9} \text{ s}$ **e** 2.3 , same

9 d $3.35 \times 10^{-9} \text{ s}$ **e** 1.005 , same **10** 1.15 s

11 The equator clock is moving faster relative to poles. It is also accelerating and hence will run slower. The effect is well below what we can detect.

6.4 Time and space

1 D **2** B, C **3** D **4 a** 1 m **b** 0.44 m **5 a** 1 m **b** 1 m

6 a $0.866c$ **b** No, $0.968c$ **7** A, C **8** A, B

9 It would represent something being in many places at the same time.

10 It is very close, even up to 10% of c . It breaks down beyond about 25% of c .

11 $\gamma = 1 + 3.6 \times 10^{-10} = 1.000\,000\,000\,36$.

6.5 Momentum, energy and $E = mc^2$

1 We see the ship foreshorten and their time slow right down. **2** C, D

3 A & B are both valid ways of looking at it.

4 a Only very slightly **b** Yes, they experience normal acceleration ($a = F/m$).

5 If the ship has a high speed in our frame, we see only a small acceleration. In the frame of the ship, the

acceleration is normal.

6 As its speed approaches c , its mass increases and so the acceleration decreases. **7** D **8 a** 2 g **b** 2 g

9 $\sim 4000 \text{ years}$ **10** True, but it is a very sound theory and any replacement will probably expand on it, not replace it.

Chapter review

1 D **2** B **3** A **4** A & C **5** B **6** C **7** C **8** C & D

9 A, B, C **10** A & C **11** B **12** C **13** D **14** C **15** B **16** A

17 A, well established by Einstein's time, B, basis of all modern physics, C, important, D, did none

18 A, true, B, false, C, true, D, false.

19 Aristotle's ideas agreed with everyday observation.

In space, we see constant velocity

20 Light had a constant speed. In conflict with Galilean relativity

21 a Light travelled with a fixed velocity relative to the aether. **b** Could not measure the speed directly accurately enough. **c** No difference in the speed of light. **22 a** 128 min **b** 139 min **c** 150 min .

23 M-M looking for similar result. Equivalent to the pilot finding the wind has no effect on the plane's motion!

24 No difference between (i) and (iii); in (ii) object will tilt.

25 Space and time are interdependent.

26 Normal for you, slow for Mars. **27 a** 5.6 years

b 2.4 years **c** Apparent distance travelled much less than 5 l.y.

28 a About 1.4 mm **b** No

29 a This is γ with $v/c = 0.995$ **b** No

c About 25.1 years **d** 2.51 years **e** No **f** No time

30 a 1.7 mm **b** No! **c** Electrons have much greater mass

31 a No, went with energy. **b** $9 \times 10^7 \text{ GJ}$.

c About 70 days . **d** One-third of a gram

32 a $4.2 \times 10^{-12} \text{ J}$ **b** 9.2×10^{37} every second

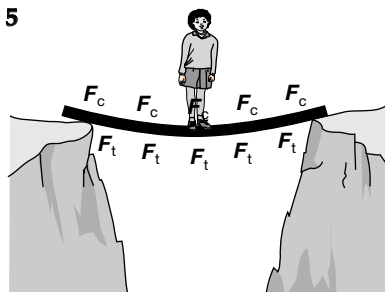
c $370 \text{ billion tonnes}$ **d** Mass associated with energy radiated into space

Chapter 7

7.1 External forces acting on materials

1 B **2** D **3** A **4 a** The steel towers are under compression. **b** The wind pushes against the blades at the top of the tower and the ground pushes against the base of the tower.

5

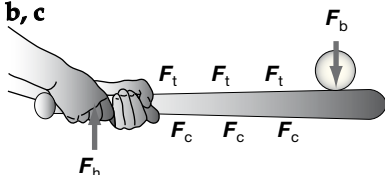


6 a 29 N b length is increased

7 a tensile b compressive

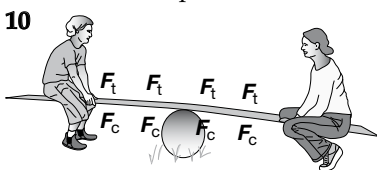
8 a compressive

b, c



9 a 780 N b rope is stretched

10



7.2 Stress and strength

1 a $7.5 \times 10^6 \text{ Pa}$ b $6.5 \times 10^5 \text{ Pa}$ 2 a $8.8 \times 10^3 \text{ N}$ b 19 MPa 3 $1.27 \times 10^8 \text{ N m}^{-2}$ 4 39.3 kN 5 7.9 kN6 a $5.1 \times 10^4 \text{ N}$ b $3.9 \times 10^4 \text{ N}$

7 a As the beam bends, the material here is not under stress—this is a neutral zone. b As the beam bends, the bottom surface is under tension. Steel is weaker under tension, so more material is needed here.

8 a 6.2 cm b 2.5 cm

9 In a beam that supports a load, the underneath surface of the beam becomes stretched and is therefore under tension. Since concrete is weaker under tension than compression this is where the cracks may appear.

10 a Reinforced concrete combines the properties of the high tensile strength of steel with the high compressive strength of concrete to produce a composite material that is strong under both tension and compression.

b The reinforcing should be located near the top surface of the concrete. As the roots expand, the concrete will bend upwards and the top surface will be under tension. Concrete is weak under tension and so reinforcing is needed here.

7.3 Strain

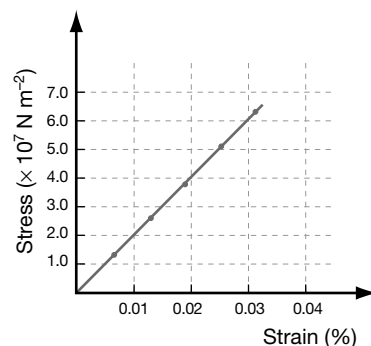
1 a equal b thin c thin 2 6.35×10^{-4} 3 5.12 mm

4 3.15 mm 5 8.006 cm 6 0.100% 7 2.8 mm

8 a

F (kN)	ΔL (mm)	σ ($\times 10^7 \text{ N m}^{-2}$)	ϵ (%)
1.0	0.013	1.27	0.0065
2.0	0.025	2.55	0.0125
3.0	0.038	3.82	0.0190
4.0	0.051	5.09	0.0255
5.0	0.063	6.37	0.0315

b



c $2.0 \times 10^{11} \text{ N m}^{-2}$ d When steel is subjected to different loads the ratio stress/strain is constant so $\sigma = k\epsilon$ where k is a constant of proportionality.

9 0.018% 10 0.0012 or 0.12%

7.4 Young's modulus

1 a $4.0 \times 10^8 \text{ N m}^{-2}$ b 2.0×10^{-3} c $2.0 \times 10^{11} \text{ N m}^{-2}$ d Young's modulus e $6.0 \times 10^8 \text{ N m}^{-2}$

2 a 5.0 mm b 2.0 mm

3 a 1.6 mm b 0

4 a ductile b 3.0 mm c No, since the material no longer behaves elastically at this tension.

5 $1.25 \times 10^{11} \text{ N m}^{-2}$

6 a The higher the value of Young's modulus, the stiffer the material. b tungsten, steel, aluminium

7 $5.5 \times 10^4 \text{ N}$ 8 a 2.5 mm b Tension; their tensile strength is lower than their compressive strength.

c As we age, calcium loss causes the cross-sectional area of our bones to decrease. For any given load, therefore, the stress the bone will experience will be greater.

9 $1.2 \times 10^7 \text{ N m}^{-2}$ 10 a C, D b A c D d B e A f A

7.5 Strain energy and toughness

1 a $4.0 \times 10^5 \text{ J m}^{-3}$ b 16 J c It is in the elastic region of the graph, so will resume its original length and no energy will be wasted as heat.

2 a $\sim 1.35 \times 10^6 \text{ J m}^{-3}$ b $\sim 54 \text{ J}$ c It is past the elastic limit, so the alloy will be permanently stretched and will heat up as the stress is removed.

3 a It breaks b $\sim 1.9 \times 10^6 \text{ J m}^{-3}$ c $\sim 76 \text{ J}$

4 In elastic deformation, all the strain energy is returned by the material as it regains its initial

dimensions after the strain is removed. For plastic deformation, the strain energy will be converted into heat energy in altering the atomic structure of the material. Consequently, the material never regains its initial dimensions.

5 a $9.55 \times 10^7 \text{ N m}^{-2}$ **b** $2.28 \times 10^4 \text{ J m}^{-3}$ **c** 0.143 J

6 No, the rod fails at a point past the elastic limit where the behaviour of the material is non-linear. A stress-strain graph is needed.

7 **1.5** **8 a** $9.5 \times 10^5 \text{ J m}^{-3}$ **b** $6.5 \times 10^5 \text{ J m}^{-3}$

9 Material P is tougher because it can absorb a greater amount of strain energy per unit volume before failing.

10 a Material P—it has a greater value of Young's modulus. **b** Material P—it experiences a greater stress value prior to failing.

7.6 Force in balance: translational equilibrium

1 A, B, D **2 a** 400–1000 N **b** 7000–12000 N

c 100 N **3** 485 N **4** $2.9 \times 10^4 \text{ N}$ upwards

5 $9.8 \times 10^5 \text{ N}$ upwards **6** 32 N **7** 5.2 kg **8** 5.0 kN

9 a $F_A = 1.13 \times 10^3 \text{ N}$, $F_B = 565 \text{ N}$ **b** cable A

c $\sigma(A) = 750 \text{ MPa}$, $\sigma(B) = 380 \text{ MPa}$ **d** No

10 Both cables would fail.

7.7 Torque

1 a spindle; 3 cm **b** front wheel axle; 1 m **c** the end of the tweezers; usually a few centimetres. **d** where the screwdriver is in contact with the rim of tin; 15–30 cm

2 a Length to lever arm is increased. **b** Force can be applied at a large distance from the pivot.

3 880 N m **4** 250 N

5 a $4.9 \times 10^5 \text{ N m}$ **b** Crane will have a counterweight providing an opposite torque.

6 a 4.9 N m **b** 9.8 N m **c** 4.9 N m

7 Provides a large counter-torque should the performer overbalance. Only a small movement of the pole is need to balance the torque produced by the performer's body when they overbalance.

8 No. The centre of mass lies outside the base of support. The bench or supports should be moved so that the centre of gravity is between the supports. Otherwise, he could use bolts to attach the beam to the left-hand support.

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 your torso to the right. You compensate for this by leaning to the left or extending your left arm. **b** $\sim 70 \text{ N m}$

10 a $3.4 \times 10^4 \text{ N}$ **b** The effective lever arm remains at 15 m throughout, so torque does not change.

c $5.2 \times 10^5 \text{ N m}$ clockwise about the pivot

7.8 Structures in translational and rotational equilibrium

1 16 kg **2** The adult should sit near the pivot; the child should sit at the end of beam. **3** 29 N m **4 a** $\tau = Ft \times 10 \sin 30^\circ = 5 \times Ft$ **b** $Ft(v) = Ftsin30^\circ$, $Ft(v) = Ftcos30^\circ$ **c** $3.4 \times 10^3 \text{ N}$ **5** 3.2 kN up and 1.7 kN up **6** 2.5 m from end X or 17.5 m from end Y **7 a** 160 N m **b** 250 N m **c** 750 N m **8 a** $F_x + F_y = 900 \times 9.80 = 8820 \text{ N}$ **b** $(F_y \times 1.8) - (8820 \times 2.0) = 0$ **c** 9.8 kN up **d** $-(F_x \times 1.8) + (8820 \times 0.2) = 0$ **e** 980 N down, tension **9 a** $F_r = 300 \text{ N}$ up, $F_l = 580 \text{ N}$ up **b** 1.7 m from left-hand end **10 a** (i) $F_A = 980 \text{ N}$ down, $F_B = 4.9 \text{ kN}$ up, $F_C = 4.9 \text{ kN}$ up, $F_D = 980 \text{ N}$ down (ii) $F_A = 1.5 \text{ kN}$ down, $F_B = 5.9 \text{ kN}$ up, $F_C = 5.9 \text{ kN}$ up, $F_D = 1.5 \text{ kN}$ down **b** As the woman walks from A to B, the force acting in pillar A decreases and the force acting in B increases. When she passes point B and continues on to point P, the forces in both A and B increase, in order to produce a greater torque to counterbalance the increase in torque as she moves to point P. **11** 52 N **12 a** $F_h = 390 \text{ N}$, $F_v = 680 \text{ N}$ **b** 330 N

Chapter review

1 D **2** C **3** C **4** D **5** E **6** B **7** E **8** A **9** B **10** B

11 a $1.3 \times 10^{11} \text{ Pa}$ **b** 0.25% **c** Young's modulus depends only on the properties of the material itself, whereas the spring constant depends on the nature of the material as well as its length and thickness.

12 a 12.5 cm **b** 400 J **c** D **d** The concrete is poured over a lattice of steel rods and sets around them.

13 $3.0 \times 10^8 \text{ Pa}$ **14 a** 0.16 mm **b** $3.0 \times 10^4 \text{ N}$ **c** 4.0 kN

15 6.6 J **16 a** Stiffness is the resistance of the material to being stretched or compressed. **b** nylon, aluminium, steel **c** steel 0.025 J m^{-3} ; aluminium 0.071 J m^{-3} ; nylon 7.1 J m^{-3} **d** The area under the graph up to the point of fracture is greater for steel than it is for aluminium. **17** 18 N, 23 N

18 $F_l = 150 \text{ N}$, $F_r = 154 \text{ N}$

19 a 1.9 m **b** This reduces the torque acting on the crane, making it less likely to topple over.

20 a 250 N m^{-1} **b** Y, the force will cause the barrier to bend slightly causing compression at X and tension at Y. Concrete is weaker under tension, so the barrier is more likely to crack at Y.

Chapter 8

8.1 Principles and practicalities of electronic design

1 a 2.0 A **b** Towards the junction

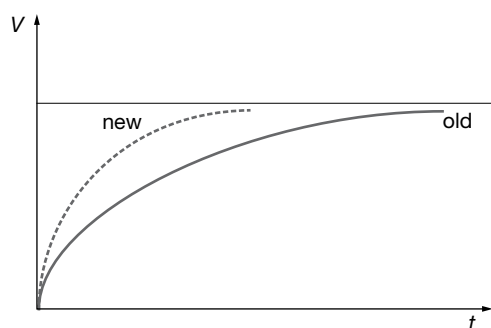
2 a (i) D (ii) A (iii) B (iv) C **b** (i) A (ii) D (iii) D

3 a 4 V **b** 5 V **c** 2.3 mA **d** 0.5 V **e** 0.5 V

4 a $2000\ \Omega \pm 10\%$ **b** $530\ 000\ \Omega \pm 5\%$ **c** brown, green, green and gold **d** grey, red, red and silver
5 a A **b** C **c** B **d** For all three cases, $R = 20.0\ \Omega$
6 a LED **b** power diode **c** zener diode
7 a (i) 10 V **(ii)** 3.5 V **(iii)** 500 Hz **b (i)** $10\ \text{V cm}^{-1}$ **(ii)** $50 \times 10^{-6}\ \text{s cm}^{-1}$ **c** gain control = $20\ \text{V cm}^{-1}$, time-base off **8 a** $10\ \Omega$ **b** 6.0 V **c** V_{out} increases **d** No. The simplest solution would be to swap the positions of the LDR and the $100\ \Omega$ resistor.
9 a 250 Hz **b** 50 V **c** 35 V **d** $33\ \Omega$; no, much lower
10 a 75 W **b** 37.5 W **c** The RMS value gives a more accurate idea of the average power, as the peak power only occurs instantaneously twice in each cycle.

8.2 Capacitors and time-varying circuits

1 C **2 a** 50 C **b** 4 mC **c** 80 pF **d** 0.1 nF **e** 20 kV
3 a The time required for the capacitor to charge up to 63% of its full charge. It is found from the product $R \times C$ in a charging circuit which includes a resistor and a capacitor. **b (i)** $1.00 \times 10^{-8}\ \text{s}$ **(ii)** 0.40 s **(iii)** $2.00 \times 10^{-7}\ \text{s}$ **4 C** **5 a** 2.0 s **b** 10.0 s **c** 0.0 V **d** 3.7 V **6 a** 1.6 s **b** $53\ \mu\text{F}$ **c** 1.7 mC **7 A**
8 a 25.0 s **b** 12.6 V **c** $29.6\ \mu\text{A}$ **d** C
9 a 5.0 s **b** 7.4 V **c** C
10 a By reducing the resistor to half of the initial value the time constant will also be halved, as $\tau = R \times C$. Thus the capacitor will be fully charged in half of the original time, and the new graph will be steeper and will look like this:

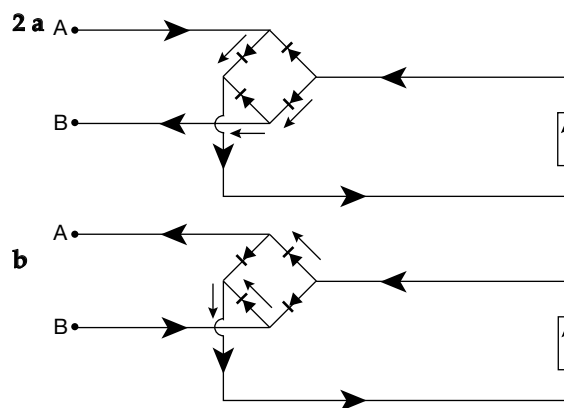


b

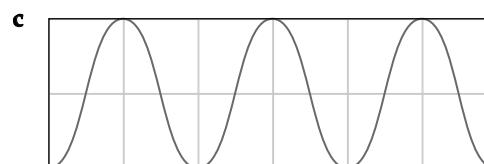
t	$t/2$	$e^{-t/2}$	$V = 10e^{-t/2}$
0.0	0.0	1.0	10.0
1.0	0.5	0.607	6.07
2.0	1.0	0.368	3.68
3.0	1.5	0.223	2.23
4.0	2.0	0.135	1.35
6.0	3.0	0.050	0.50
8.0	4.0	0.018	0.18
10.0	5.0	0.007	0.07

8.3 Rectification and power supplies

1 a To allow half of each AC cycle through so that only unidirectional current is obtained. **b** Half-wave rectifier **c (i)** A **(ii)** A **(iii)** C

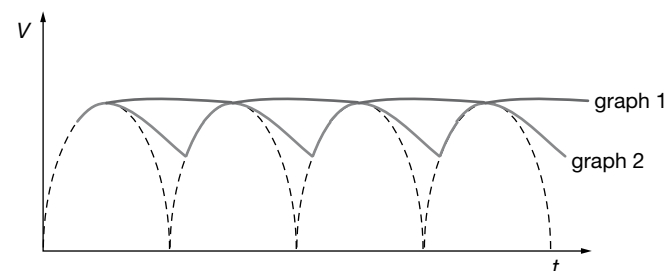


3 a During the upward pulses: 0.05 to 0.15 s, 0.25 to 0.35 s, and 0.45 to 0.55 s. **b** When the CRO shows 0 V: 0.0 to 0.05 s, 0.15 to 0.25 s, 0.35 to 0.45 s, and 0.55 to 0.60 s.



d 5.0 Hz **4 a** A **b** B

5 a With the large capacitor added the voltage across the resistor will be almost flat with a little ripple (graph 1):



b A smaller capacitor will have a smaller time constant and hence more ripple (graph 2). **c** A capacitor is used to smooth the output, to provide current between the pulses from the rectifier. The larger the capacitor, the smoother the output and the closer it comes to pure DC.
6 a At the maximum load current of 1.0 A there will be around 50% ripple voltage. At the minimum load (1/100th the current) there will be very little ripple.
b The charge stored in the capacitor will be 0.012 C. A current of 10 mA = 0.01 A would last 1.2 s. A constant current of 1.0 A would last only 0.012 s.

c The calculation in part b assumes a steady current as the capacitor is drained of charge. But as the capacitor discharges, the voltage drops quite rapidly and thus less current will flow to the load, causing a ripple in the output voltage. **d** Increase the size of the capacitor or use a zener diode or a voltage regulator circuit.

7 a $15\ \Omega$ **b** 0.45 W, 1.95 W wasted **c** The radio would work if a lower value for the dropping resistor was used (but with more power wasted) but not if a higher value was used.

8 The zener diode and dropping resistor will both carry current, which will result in some loss of the power provided by the supply. Also the value of the dropping resistor must be carefully chosen. If its value is too high, the voltage across it will fall too far when the current increases. If the value is too low, the current through the zener diode may exceed its power rating and cause it to fail. **9** B, D

10 a To smooth the output of the bridge rectifier to around 8 V **b** To keep the input to the regulator above 7 V **c** The extent to which the heat produced can be dissipated. If the regulator gets too hot it shuts down until it cools off. Without a heat sink this regulator can only dissipate 3 W which means it will have a maximum output current of about 1 A.

d The difference between the input and output voltages would be 20 V, which means 20 W must be dissipated, and wasted, for every 1 A of current. This would limit the current to 1 A even with the best heat sink (as maximum power dissipation is 20 W). **e** 3.5 W

8.4 Constructing and testing a working power supply

1 All of them **2 a** If a diode breaks down and allows a reverse current to flow, it effectively becomes a short circuit and will allow a high current to flow directly from one side of the transformer (through another diode) to the other every half cycle. **b** The short-circuited diode is effectively preventing a potential across the load during half of the cycle.

3 For most capacitors it does not matter, but electrolytic capacitors must be placed so that the lead labelled negative is towards the negative side of the circuit. This is because the chemical action that is used to create the insulating layer can be reversed if current passes through the wrong way; this can have explosive results.

4 a 0.25 W or more **b** 3.3 Ω , 7.5 W or more **c** $I = 0.5\text{ A}$, power 2.5 W; this is over twice the rated power of the resistor and so the resistor will very likely burn out.

5 a A, D **b** C **c** B

6 a The diode on the right should not be connected to the top line. **b** This diode will short the supply on

each negative cycle and probably burn out.

c The top end of the diode should be connected to the $+V_{\text{out}}$ line (row 9).

7 a C **b** B **c** The time constant = 1 s for this capacitor. This is 100 times the period and will give a good smooth output. The larger value capacitor would be bulky and more expensive.

8 6 V. The capacitor will charge to the peak output voltage from the transformer which is 7 V, but there will be about 1 V drop across the diodes.

9 The regulator gives a constant output provided it is supplied with a voltage around 2 V (or more) higher. Fluctuations above this are removed. Thus a moderate amount of ripple can be eliminated.

10 a Current flowing through the regulator must undergo a voltage drop, ΔV , from input to output. This results in a power loss of $P = \Delta V \times I$. **b** This power may overheat the regulator and cause it to shut down. **c** Regulators are often mounted on a heat sink to remove this heat as quickly as possible.

Chapter review

1 B **2** D **3** C **4** A **5** B **6** D **7** C **8** A **9** D
10 A **11** B **12** A **13** B **14** B **15** C **16** B, D **17** D
18 D, C **19** B, C **20** A **21 a** 125 Ω **b** 50 mA **c** 120 Ω
22 a 20 **b** 0.2 mC **c** 20 μF **d** Approx. 10 ms **e** 500 Ω
f 100 Ω

23 Expected current 0.454 mA which is >1% of actual. Due to resistance of meter. 0.5 V across meter, so R of meter 1.2 k Ω .

24 a 200 Ω . Zero current through the diode
b 15 mA **c** 12 mA

25 Excessive current in other diodes or transformer—very hot or burn out. A fuse can protect. Heat sensor can shut off supply.

26 A particular 7805 will have a constant V_{out} within 4.8–5.2 V. V_{out} will be constant to 3 mV with V_{in} in range 7–25 V.

27 a 21 V **b** 20 V **c** The ripple would represent about 20% of the peak voltage. **d** The input voltage to the regulator will vary between 20 V and about 16 V. These figures are below the maximum and above the minimum and so the regulator should operate satisfactorily. **e** Diodes 1.5 W; regulator ~12 W; load 18 W **f** Yes; it will require a moderate heat sink.

Chapter 9

9.1 Fundamentals of magnetism

1 A magnetic field exists at any point in space where a magnet or magnetic material (e.g. iron, nickel, cobalt) will experience a magnetic force. **2** C **3** B **4** A **5** C

6 While in a magnetic field a magnet will be subjected to forces which produce a torque that will tend to align the magnet with the field.

7 a north **b** north-east **c** east

8 a east **b** west **9 0** **10 a** east

b north **c** 30° east of north **11 5B**

9.2 The foundations of electromagnetism

1 A east, **B** south, **C** west, **D** north

2 A west, **B** north, **C** east, **D** south **3** east **4** west

5 a south **b** north **c** Resultant magnetic field due to the conductors is zero. The only field in existence at that point would be that due to the Earth, whose direction is north.

6 R. Only if the S field from m and n is balanced by the Earth's N field.

7 a south **b** south **c** south (assuming it is greater than the Earth's field) **8 a** west **b** east

9.3 Currents, forces and fields

1 a 0.40 N up **b** 0.25 N down **2** 1000 : 1 **3 B**

4 a 2.0×10^{-4} N north **b** 1.0×10^{-4} N south

5 a 1.5×10^{-3} N west **b** 3.0×10^{-3} N east

6 a 40 A into page **b** 20 A out of page

7 a 2.0×10^{-2} N m^{-1} west **b** 2.0×10^{-2} 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** north

10 a 2.0×10^{-9} N **b 0** **c** 2.5×10^{-4} N

9.4 Magnetic fields around currents, magnets and atoms

1 Into page **2** Out of page **3** Due to the circular geometry, the field inside the loop is more concentrated than the field outside the loop. There is a greater magnetic flux density inside the loop.

4 a B into page **b 3B** into page **c 0** **5 A** **6** south

7 south **8 A, C** **9 B, C** **10 C**

9.5 Forces on moving charges

1 C 2 A 3 a south **b C**

4 A (the speed remains constant, the direction changes)

5 a north **b A**

6 A particle that does not carry an electric charge, e.g. a neutron. **7 C, D** **8 A, B**

9 a south **b (i) 2F (ii) 2F (iii) 4F**

10 a 2F, north **b** Very much less curved (i.e. greater radius)

9.6 Electric motors

1 1.0×10^{-2} N, into the page

2 1.0×10^{-2} N, out of the page **3 0** **4** anticlockwise

5 D 6 a down **b** up **7** anticlockwise

8 a down **b** up **c 0 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 a Into page **b** Out of page **c** Out of page

d Out of page **2** 5.0×10^{-5} T south **3** Into page

4 1.0×10^{-4} T north **5** 7.1×10^{-5} T **6** north-west **7 C**

8 a 5.0×10^{-9} N into page **b** 2.0×10^{-3} N into page

c 5.0×10^{-2} N into page **9 a** Out of page **b** Into page

10 0 11 a attraction **b** attraction **c** repulsion **12 0**

13 2.0 N m^{-1} **14 B, C** **15 B** **16** Using the right-hand

rule gives direction of the field, which must be into the page for B_y . Since the amount of deflection is the same, the force on each electron beam is equal, and since magnetic fields and charges are also equal, $v_1 = v_2$ ($F = Bqv$). **17 C** **18** 0.10 N **19** anticlockwise **20 D**

Chapter 10

10.1 Magnetic flux and induced currents

1 3.2×10^{-6} Wb, 2.3×10^{-6} Wb, 1.6×10^{-6} Wb, 0

2 0.9×10^{-6} Wb, 1.6×10^{-6} Wb, 2.3×10^{-6} Wb, 3.2×10^{-6} Wb

3 a 3.2×10^{-6} Wb **b** 6.4×10^{-6} Wb **c** 3.2×10^{-6} Wb

d 1.6×10^{-6} Wb

4 a zero **b** negative **c** positive **d** negative

5 There must be a non-zero rate of change of magnetic flux.

6 A momentary negative current, zero current, a momentary positive current

7 A positive current then a negative current.

8 a 1.0×10^{-5} Wb **b 0**

9 a 1.0×10^{-5} Wb **b** 0.01 Wb s^{-1}

c 4.0 mA from Y to X

10 a 8.0 mA from X to Y

b 1.0 mA from X to Y **c** 2.0 mA from X to Y

10.2 Induced EMF: Faraday's law

1 a 1.2×10^{-6} Wb **b 0** **c** 3.0×10^{-5} V **d** 2.0×10^{-5} A

2 a 8×10^{-5} Wb **b** 4.0×10^{-3} V **c** 2.0 V

3 a 1.0×10^{-4} Wb **b** 6.0×10^{-3} V

4 a 2.5×10^{-3} Wb **b** 0.02 V **5 1:2 6 D**

7 5.0×10^{-2} T

8 a 0 **b** 8.0×10^{-3} V **c** By pulling the loop from between the poles of the magnet with greater speed.

9 a 0 **b** 2.0×10^{-3} V **c 0** **d** 1.0×10^{-3} V **10** 5.8×10^{-6} V. No, each radius is moving in an opposite direction.

10.3 Direction of EMF: Lenz's law

- 1 C 2 a** Out of page **b** Into page **c** Out of page
3 a anticlockwise **b** clockwise **c** As the centre of the magnet passes through the ring. **d** Strength of the magnet, velocity of the magnet, diameter of ring, resistance of circuit containing the ring.
4 a right **b** right **c** left
5 a E. A steady positive current through Q can only be produced by a negative current through P which is changing at a constant rate. This is because $I_Q \propto -\Delta\phi_B/\Delta t$ in P where $\Delta\phi_B/\Delta t$ in P $\propto \Delta I_P/\Delta t$ (since the amount of magnetic flux through a coil is proportional to the current in the coil). **b** $D \Rightarrow I_Q \propto -\Delta I_P/\Delta t$
c A, B, C $\Rightarrow I_Q \propto -\Delta I_P/\Delta t$
6 a $X \rightarrow Y$ **b** left **7 a** $Y \rightarrow X$ **b** left **8 A**

10.4 Electric power generation

- 1 a** 1.0×10^{-2} mWb **b** 0.97×10^{-2} mWb
c 0.87×10^{-2} mWb **d** 0.71×10^{-2} mWb
e 0.50×10^{-2} mWb
f 0.26×10^{-2} mWb **g** 0
2 0.3, 1.0, 1.6, 2.1, 2.4, 2.6 mWb s⁻¹
3 The rate of change of flux increases
4 0.03 V, 0.10 V, 0.16 V, 0.21 V, 0.24 V, 0.26 V
5 a Maximum rate of change of flux occurs at 90°.
b 0.263 V **6 B 7 a C b D c C d B e D 8** 0.81 T

10.5 Alternating voltage and current

- 1 D 2 C 3 a** 339 V **b** 679 V **c** 3.39 A **d** 2.40 A
4 D 5 a 50 Hz **b** 240 V **c** 85 V
6 a 96 Ω **b** 339 V **c** 3.53 A
7 a 2.5 A **b** 3.5 A **c** 68 V **d** 9.6 Ω
8 a 120 V **b (i)** 144 W **(ii)** 144 W **c** 120 **d** 289 W
9 a 7.1 V **b** 4.2 A **c** 30 W

10.6 Transformers

- 1 a** $V_p = N_1 \Delta\phi_B/\Delta t$ **b** $V_s = N_2 \Delta\phi_B/\Delta t$ **c** $V_p/V_s = N_1/N_2$
2 a A, B, D **b** A, B, D **3 a A b B, 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 V **b** 0.20 A **c** 16 W
7 a 40 **b** 0.14 A **c** 24 W
8 No. A direct current cannot induce an EMF in the secondary coil because $\Delta\phi_B/\Delta t = 0$.
9 There will be no power consumed in the primary circuit of an ideal transformer. This is because the induced current in the primary coil is equal in magnitude but opposite in direction to the applied current. In a real transformer like this one there may

be a loss of around, say, 2 W, which would add up to 1200 J in 10 minutes.

10.7 Using electrical energy

- 1 a** Lower current is required and so thinner cables can be used. **b** Losses such as current leakage via insulators and the atmosphere.
2 a 2.00×10^3 A **b** 1.00×10^3 A
3 a 8.0% **b** 2.0% **4** 1.0%
5 a 10 A **b** 400 W **c** 8.0% **d** 460 V
6 a 1.0 A **b** 0.08% **c** 4996 V
7 a 30c **b** 0.6c **c** 45c **d** 15.1c **e** \$3.94
8 a 2 MA, 4 MV, impossible! **b** 5 kA, 10 kV, 90 kV
c 25 MW
9 a 15 A **b** 30 V, 9970 V **c** 450 W, No **10 a** 150 A
b 300 V, 700 V **c** 45 kW, 105 kW **d** No, 30% power lost

Chapter review

- 1 a** 3.2 mA clockwise **b** 3.2 mA anticlockwise **c** 1.6 mA anticlockwise **2 a** 2.0 mV **b** 2.0 mA
3 a A b C c B 4 20 mA from Y to X
5 A, C 6 From X to Y **7 a** 4.0 mA from X to Y
b 8.0×10^{-6} N left
8 No induced current as circuit is open.
9 a 1.6 mV **b** 0 **10** 5.0×10^{-5} A
11 a 1.0 A **b** 10 **c** 42 W **12 a C b A c B**
13 a 500 Hz **b** 17.7 V **c** 50 V
14 a 188 W **b** 375 W **c** 1.67 Ω **15 D 16 D**
17 a 25.2 V, 8.9 V **b** 25.2 V, 17.8 V **c** 90 Hz **d** 9.5 V
18 a 2.0 A, 236 V, Yes **b** Too much voltage drop along cable **c** 64 V, 3.75:1 **d** 800 W
19 Appliances with transformers or AC motors would not work and could burn out. Lights and heaters would work normally, but at full load there would be a voltage drop of about 50 V along the power line which would leave only 190 V.
20 Step down, turns ratio 5 : 1, 240 V, 238.5 V, 237 V, yes.

Exam-style questions—Electric power

- 1 a A b B c G 2 a** To left (induced magnetism)
b left **c** right **3** 1.0 mN **4** west to east
5 4.9×10^4 A **6** 2.0 mN **7 B 8** 50 mN **9** right
10 10 mN **11** left **12 a (i, ii)** zero **b (i)** 0.50 N out
(ii) 0.5 N in **c** Rotate 90° **13 A, B, C**
14 a AB up, CD down **b** Position shown **c** Vertical, motor continues to turn due to its momentum which propels the coil past this point of zero torque.
15 a 4.0 A **b** 20 N **c** 64 N
16 Direction of induced current is from Y to X. The direction is predicted by the right-hand force rule. Direction of the magnetic force on side XY must oppose

the entry of the loop into the field. **17** 4.0 mV
18 3.2×10^{-5} W **19** The external force that is moving the loop towards the cube **20** 6.4×10^{-4} N
21 0. After 1.2 s the entire loop is moving in a region of uniform magnetic field. This means that the rate of change of flux and hence the induced EMF are zero.
22 Direction of induced current is from X to Y. The direction is predicted by the right-hand rule. Direction of magnetic force on the side of the loop still in the field must oppose its exit from the field.
23 a 5.0 μ Wb **b** 0 **c** 2.5 mV **d** 1.25 mA
e No, it rapidly loses energy.
24 a 200 μ A **b** 0.21 T
25 a Sine wave amplitude 0.9 V, period 0.01 s
b 0.64 V **c** Half period, twice amplitude, 1.3 V
26 314 V **27** 222 V **28** B **29** D **30** C **31** 500 Hz
32 20 V **33** 7.1 V **34** 5.0 W **35 a** 0.40 A **b** 6000 V
c 200 **d** 850 W **e** 1700 W **36** C **37** C
38 Voltage drop on line, 218 V
39 1:20 step up and 20 : 1 step down.
40 a 0.8 A **b** 1.6 V **c** 1.3 W **d** 249.9 V, 3999 W
41 13% and 0.03%, power loss = $I^2 R$

Chapter 11

11.1 Review of light and waves

1 Both involve the transfer of energy without a net transfer of matter.
2 Mechanical waves involve the physical transfer of vibration from particle to particle within the medium. Denser materials have closer particles and so this transfer occurs more readily.
3 $f = 160$ Hz, $T = 6.3 \times 10^{-3}$ s **4** 1.8 m s⁻¹ **5** 0.288 m
6 Decrease, allowing more time, and therefore more distance between wavefronts. **7** A **8** D
9 A wave can reflect part of its energy at the glass (creating the image of yourself) and allow part of its energy to continue through the glass to be reflected from the items within. Particles could not split themselves up.

11.2 The wave model established

1 a A source that produces plane light waves that are in phase. **b** The plane waves will diffract, and emerge from the slits as two sets of circular wavefronts.
c If a crest is emerging from one slit, then the wavefront simultaneously emerging from the other slit will also be a crest.
2 a $p_d = 0$, i.e. constructive interference will occur at this point. **b** When a crest and a trough meet, producing destructive interference
c (i) dark (ii) bright

3 a dark **b** bright **c** dark
4 a The fringes are closer together and brighter.
b Since fringe spacing increased, the wavelength must have increased. Therefore, frequency was decreased.
5 A, C **6** B
7 a interference **b** The central band will be wider than the others and brighter.
8 a Laser light is best as monochromatic coherent light will result in constructive and destructive interference effects in a regular pattern of bright and dark fringes
b B, D
9 For appreciable diffraction to occur and create a blurred image, the approximate size of the diffracting object would be equal to the wavelength of visible light, i.e. around 10^{-7} m.

11.3 Photoelectric effect: Counterevidence for wave model

1 a 5.4×10^{-7} m **b** 3.7×10^{-19} J, 2.3 eV **2** For each metal there is a threshold (minimum) frequency which must be used for electron release to occur; the rate of electron release is proportional to the light intensity; there is no time delay in the release of electrons.
3 All statements are true.
4 a 2.30 eV **b** 4.24×10^{-19} J
5 a (i) 1.95 eV (ii) 3.12×10^{-19} J **b** 8.28×10^5 m s⁻¹
6 a 2.0×10^{16} Hz **b** 1.33×10^{-17} J, 83 eV
7 The same as Figure 11.24a. The brighter light results in a greater photocurrent but the same stopping voltage since the metal cathode and the incident light frequency are unaltered.
8 1×10^6 **9** 9.0×10^{20} **10** 1.6×10^{15} Hz

11.4 The dual nature of light

1 C **2** B, D **3 a** false **b** true **c** false **d** true **4** E
5 B, E **6** 4.3×10^{-15} eV s, 2.2 eV
7 a 4.6×10^{14} Hz **b** 1.9 eV **c** 2.0 eV **d** 0.10 eV
e 1.7×10^{-25} kg m s⁻¹
8 a -0.10 V **b** The photoelectrons do not have sufficient kinetic energy to reach the anode. **c** -0.25 V

Chapter review

1 The particle model predicted that as light sped up it would be refracted towards the normal, but it is refracted towards the normal as it slows down as the wave model predicts.
2 The central band is white since constructive interference occurs here for all component wavelengths of the white light. Since the degree of diffraction is dependent upon wavelength, the different colours will produce bands at different locations, resulting in the

side bands having coloured edges and merging into one another.

3 $3.1 \times 10^{-11} \text{ m}$ **4** The frequency is unaltered and the wavelength will be 50% longer.

5 a A series of alternate bright and dark fringes.

b (i) width of central bright fringe will decrease

(ii) width of central bright fringe will increase

6 a Fringe spacing will increase. **b** Fringe spacing will decrease. **c** Fringe spacing will remain unchanged.

d Fringe spacing will remain unchanged.

7 a (i) $3.43 \times 10^{-19} \text{ J}$ **(ii)** 2.14 eV

b $1.14 \times 10^{-27} \text{ kg m s}^{-1}$ **c** 1.46×10^{21}

8 a $4.60 \times 10^{-19} \text{ J}$ **b** $1.53 \times 10^{-27} \text{ kg m s}^{-1}$ **c** 5.00×10^{20}

d 2.30 mW **9 a**

10 a The wave model predicts that light of any frequency will emit photoelectrons from a metallic surface. **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 low intensity beams of light will take longer to eject photoelectrons from a surface. **c** According to the wave model of light, a higher intensity beam will deliver more energy to the electrons and consequently emit photoelectrons with a greater kinetic energy.

11 a $6.6 \times 10^{-34} \text{ J s}$ **b** Planck's constant **c** $5.0 \times 10^{14} \text{ Hz}$

d No. The frequency of red light is below the threshold frequency for rubidium.

12 a 2.07 eV **b** $4.0 \times 10^{-20} \text{ J}$ **c** $2.7 \times 10^{-25} \text{ kg m s}^{-1}$

d 0.25 V

Chapter 12

12.1 Matter waves

1 B **2 a** $7.28 \times 10^{-11} \text{ m}$ **b** $3.97 \times 10^{-13} \text{ m}$

c $6.62 \times 10^{-12} \text{ m}$ **d** $1.66 \times 10^{-34} \text{ m}$

3 a $1.7 \times 10^{-35} \text{ m}$ **b** There is no slit small enough for such a wave to pass through in order to be diffracted.

This wavelength is many times smaller than the radius of an atom. **c** No. The objects that we encounter in our daily lives are simply too massive to have a momentum small enough to produce a detectable matter wave.

4 a $1.32 \times 10^5 \text{ m s}^{-1}$ **b** 0.69 m s^{-1} **c** $1.09 \times 10^3 \text{ m s}^{-1}$

5 a $8.0 \times 10^{-18} \text{ J}$ **b** $3.82 \times 10^{-24} \text{ kg m s}^{-1}$ **c** $1.74 \times 10^{-10} \text{ m}$

6 $\lambda = h / (2mq\Delta V)^{\frac{1}{2}}$

7 C, D. The de Broglie wavelength of the 1.0 keV electrons is $3.9 \times 10^{-11} \text{ m}$. This is smaller than the interatomic spacing, but of the same order of magnitude. The other alternatives do not satisfy this criteria. **8 a** 0.38 V **b** $2.0 \times 10^{-4} \text{ V}$ **c** $2.57 \times 10^{-5} \text{ V}$

12.2 Photons shed light on atom structure

1 a The term quantisation refers to the fact that the energy levels in an atom cannot assume a continuous range of values but are restricted to certain discrete values, i.e. the levels are quantised. **b** The ground state is the lowest energy state that an atom can exist in and represents the stable state of the atom before any external energy has been added. **c** Excited states are the possible energy levels that an atom can move to after it has absorbed energy from an external source. **d** The ionisation energy is the least amount of energy required to eject an electron from an atom, and represents the highest energy state of the atom.

2 a $1.50 \times 10^{-6} \text{ m}$ **b D** **3 a** $2.46 \times 10^{15} \text{ Hz}$

b $1.03 \times 10^{-7} \text{ m}$ **c** 13.6 eV

4 a $n = 3$ **b** No. The difference between the ground state and $n = 3$ is 12.09 eV . A photon will not give up a fraction of its energy to the atom. **c** The atom would be ionised and the ejected electron would be emitted with kinetic energy 0.40 eV .

5 a $n = 4$

b 0.63 eV , 2.51 eV , 1.88 eV , 12.72 eV , 12.09 eV , 10.21 eV

6 The excess energy is given to the electron in the form of kinetic energy. A freed electron may have kinetic energy of any value, i.e. it is not quantised.

7 Bohr's model states that electrons may only have specific energy values. The energy quanta absorbed or emitted by an atom correspond to the differences between the allowed energy states.

8 $3.89 \times 10^{-7} \text{ m}$ **9** 10

10 Like all atoms, sodium atoms can emit more frequencies of light than they can absorb as the electrons can fall down through orbits in stages or in a single fall. Atom excitation occurs in single jumps from the ground state only.

12.3 Bohr, de Broglie and standing waves

1 The fall from $E = 0$ to $E = -3.4 \text{ eV}$.

2 $1.2 \times 10^{-7} \text{ m}$, $1.0 \times 10^{-7} \text{ m}$, $9.7 \times 10^{-8} \text{ m}$

3 No, higher energy end, approximately 365 nm , is not visible. **4** $1.2 \times 10^{-7} \text{ m}$

5 De Broglie proposed a model of the atom in which electrons were viewed as matter waves with resonant wavelengths.

6 The circumference of an electron orbit must correspond to an integer multiple of the de Broglie wavelength of the electron, just as the violin string can only support waves of resonant wavelengths that are integer multiples of the string's fundamental wavelength.

7 a $E_5 = -0.544 \text{ eV}$, $E_{10} = -0.136 \text{ eV}$, $E_{15} = -0.060 \text{ eV}$, $E_{20} = -0.034 \text{ eV}$ **b** The differences become smaller.

Chapter review

- 1 C** **2** $1.18 \times 10^{15} \text{ Hz}$, $1.62 \times 10^{15} \text{ Hz}$, $4.34 \times 10^{14} \text{ Hz}$
3 a A high density of electrons **b** A low density of electrons **c** Wavelength and wave behaviour **4 B**
5 a $7.90 \times 10^{-25} \text{ kg m s}^{-1}$ **b** $8.39 \times 10^{-10} \text{ m}$ **c** 1.39×10^{20}
6 No minimal wavelength. All light above ionisation energy may be absorbed.
7 See points on page 446
8 a Only frequencies matching the differences between energy levels can be absorbed. **b** Electron raised beyond the highest possible orbital level.
9 The photon is absorbed, the emitted photons go in all directions, therefore in the direction that the incident light shines the intensity of this photon is largely reduced, hence the black line.
10 Only the standing wave model provided a physical explanation for the existence of orbital levels.

Exam-style questions—Interactions of light and matter

- 1 a** 9.0×10^{20} **b** infrared
2 a 2.2×10^{20} **b** $1.5 \times 10^{-27} \text{ kg m s}^{-1}$
3 a 2.9×10^{21} **b** 1.0 kW **4 B** **5 C** **6 A**
7 5.01 eV
8 (i) Only certain frequencies of light will emit photoelectrons. **(ii)** There is no time difference between the emission of photoelectrons by light of different intensities. **(iii)** The maximum kinetic energy of the ejected photoelectrons is the same for different light intensities of the same frequency.
9 a $2.4 \times 10^{-9} \text{ m}$ **b** A series of bright and dark fringes. **c** Electrons travelling at high speed exhibit wave-like behaviour. **10** 164 m s^{-1}
11 a $f = 2.11 \times 10^{15} \text{ Hz}$, $\lambda = 1.41 \times 10^{-7} \text{ m}$
b $f = 1.15 \times 10^{15} \text{ Hz}$, $\lambda = 2.60 \times 10^{-7} \text{ m}$
c $f = 5.07 \times 10^{14} \text{ Hz}$, $\lambda = 5.92 \times 10^{-7} \text{ m}$
12 a $1.0 \times 10^{-17} \text{ J}$ **b** $4.8 \times 10^6 \text{ m s}^{-1}$ **c** $1.5 \times 10^{-10} \text{ m}$
13 Incident energy less than minimum energy difference between the lowest and next orbital level; therefore no orbital change, therefore no absorption.
14 a 1810 eV **b** $4.4 \times 10^{17} \text{ Hz}$ **c** $2.67 \times 10^{17} \text{ Hz}$ **d** Yes, atom would be ionised. **e** $5.9 \times 10^{-13} \text{ m}$
15 De Broglie would say the electrons had diffracted through the gaps between the crystal atoms.
16 In addition to their particle properties, by existing in their quantised energy states, they mimic the standing waves (quantised energy) formed in many physical objects. **17 B, C, D, E**

- 18** The energy that a single electron would gain after being moved through a potential of 1 V .
19 Electron absorbs energy required for release; excess energy results in extra kinetic energy of the electron.
20 The surprise that light can display both particle and wave properties was repeated when electrons (known to have particle properties) were found to have wave properties as well when very fast moving.
21 a Fringe spacing doubles. **b** Fringe spacing doubles. **c** Fringe spacing doubles. **d** White central band, fringes separate colours. **e** Wider central band.
22 a $1.2 \times 10^{-11} \text{ m}$ **b** Resolution affected by diffraction. Significant diffraction occurs if object size is similar to wavelength. Electrons can have much smaller wavelengths than visible light.
23 Protons would have smaller wavelength giving greater resolution, but proton microscopes are not as easy to build!
24 C **25 a** 2.0 keV **b** $3.00 \times 10^8 \text{ m s}^{-1}$
26 a Photon energy $>$ ionisation energy i.e. there is enough energy to free the electron.
b $0.4 \text{ eV} = 6.4 \times 10^{-20} \text{ J}$ **c** $3.4 \times 10^{-25} \text{ kg m s}^{-1}$
d $1.9 \times 10^{-9} \text{ m}$
27 Since there is no energy level 10.0 eV above the ground state, the photon cannot be absorbed. **28 D**
29 a 8.8 eV **b** $1.4 \times 10^{-7} \text{ m}$ **c** 10.40 eV
30 a $1.12 \times 10^{-18} \text{ J}$ **b** Since the photon energy does not match any of the differences, it will not be absorbed and the atom will not be excited
c 1.80 eV , 6.70 eV , 4.90 eV **d** $1.9 \times 10^{-7} \text{ m}$
31 $3.0 \times 10^{-8} \text{ m}$
32 a Incident energy insufficient for any excitation. At least 4.90 eV required. **b** Electrons have escaped mercury atoms and conduct current across tube.
c $2.5 \times 10^{-7} \text{ m}$ **33 3** **34** An 8.0 eV photon will not cause excitation because it does not match any energy level differences.
35 Extremely unlikely. Electron only remains in an excited state for less than a millionth of a second.
36 $2.2 \times 10^{-22} \text{ kg m s}^{-1}$
37 They must have equivalent wavelengths
38 $3.6 \times 10^{-11} \text{ m}$ **39** $3.6 \times 10^{-11} \text{ m}$
40 $1.8 \times 10^{-23} \text{ kg m s}^{-1}$

Chapter 13

13.1 Particle accelerators

- 1 B** **2 a** Electrons leave the hot cathode of the evacuated tube and accelerate towards a positively charged anode. The electrons can be deflected as they pass through an electric field produced by a pair of 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 by a high potential difference between the cathode and positively charged anode. **3** The standing-wave linear accelerator consists of a large number of drift tubes, each separated 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 and is timed such that electrons 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. This type of standing-wave linac is useful for low energy ion accelerators (less than 200 MeV) and as a result is not employed in the Australian Synchrotron, which utilizes a traveling-wave type of linac.

4 A circular accelerator, such as a cyclotron, can be used to accelerate particles within a more compact space than the equivalent operation of a very long linear accelerator. **5 a** $5.9 \times 10^7 \text{ m s}^{-1}$ **b** $2.2 \times 10^{-4} \text{ m}$

6 a The electron will experience a force at right angles to its motion. This acts upwards in the initial moment and curves the electron in an upwards arc from its starting position. **b** The radius of the electron path is dependent upon its velocity and the magnitude of the magnetic field that is acting.

7 a $1.4 \times 10^4 \text{ V m}^{-1}$ **b** $9.3 \times 10^6 \text{ m s}^{-1}$

8 $3.0 \times 10^7 \text{ m s}^{-1}$ **9** $9.4 \times 10^{-4} \text{ T}$

10 a $9.6 \times 10^{-15} \text{ N}$ **b** $4.6 \times 10^{-3} \text{ m}$

13.2 Synchrotrons

1a (i) The 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 and then accelerate across a 100keV potential difference. The electron beam travels through an ultra-high vacuum within the linac, to prevent energy loss through the interaction with air particles. As the electrons travel, focusing elements act on the beam to ensure it doesn't collide with the walls of the vacuum tube. RF (radio-frequency) cavities throughout the linac produce intense electromagnetic radiation of several hundred megahertz perpendicular to the electron beam. The RF radiation propagates through the linac as a traveling wave. Electrons are timed in pulses so that they travel through the linac in bunches which are accelerated by the RF radiation. As a result, electrons are accelerated to close to the speed of light throughout their journey

through the linac. **(ii)** Each time the charged particles travel around the circular booster ring they receive an additional energy burst from a radio-frequency (RF) chamber which the electrons pass through each time they orbit the ring. **(iii)** In the storage ring of the synchrotron, electrons revolve around the storage ring for hours at a time at speeds near that of light. A series of magnets make them bend in arcs as they travel through the ring. It is as 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 used in the circular booster ring is periodically increased as the velocity of the electrons increases to account for energy losses due to the increasing effects of relativity due to this 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, booster and storage rings. **(iv)** Insertion devices are located in the straight sections of the storage ring.

2 a The precise configuration of 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 transformers.

b The particles would gradually lose energy through collisions with other atoms and the production of synchrotron light. Their orbit speed would be slowed, they would cease to produce synchrotron light and would eventually stop. **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 magnetic poles aligned closely together in rows. The effect of the undulator is to produce much brighter synchrotron radiation. The difference in output compared with that from a wiggler is that the radiation is enhanced at specific wavelengths and is not of a continuous nature.

7 a 3 GeV **b** Any three of a wide range of industries could be listed, including pharmaceutical development, mining and exploration and manufacture of microstructure products. **c** Student response

8 a $2.1 \times 10^8 \text{ m s}^{-1}$ **b** For electrons being accelerated

across a potential difference of 2.5 keV, the effects of relativity will come into play. The effective mass of the electrons will be greater than their rest mass. As a result, they will not reach the velocity calculated in part a. (For such a potential difference, the electrons will reach about 80 % of the calculated velocity).

9 a 1.1×10^{-3} m (or 1.1 mm) **b** At such velocities so close to the speed of light, the effects of relativity have a huge impact on the radius of the electron beam. Because the mass of the electrons is about 6000 times greater than their rest mass, it follows that the expected path radius will also be some 6000 times greater than predicted in part a. Because the bending magnets are only found in sections of the storage ring, the actual path radius is even greater still.

10 Similarities could include: the same basic design, i.e. linac, booster, storage ring and a number of beamlines, a third-generation source. Differences could include: the Delta facility has a power output of 1.5 GeV compared with the proposed 3 GeV of the Australian Synchrotron, the Australian Synchrotron has a larger number of beamlines, the Delta is about half the size of the Australian Synchrotron, Delta is oval in shape, whereas the Australian Synchrotron has a storage ring that is symmetrical through all axes.

13.3 Synchrotron radiation

1 a Synchrotron light is produced when high-energy charged particles travel in a curved path. **b** High intensity, broad spectral range, being highly collimated, being highly polarised, emitted in very short pulses.

2 a Approx. 10^{-3} to 10^5 eV **b** 10^{-1} to 10^{-10} m **c** It covers the scale of cells, viruses and individual atoms, thus making synchrotron light a suitable tool to investigate these structures.

3 a 50 keV to 100 eV **b** 100 keV—hard X-rays, 1 eV—at the boundary between infrared and visible light **4 B**

5 X-rays from an X-ray tube are produced in a single burst; those generated in a synchrotron source can last for hours. X-rays from an X-ray tube pass through lighter atoms whereas synchrotron X-rays are more likely to interact with these. Synchrotron X-rays are 100 million times brighter than X-rays from traditional sources. **6 B** **7** 1.9×10^{-9} m

8 A collimated beam of X-rays produced in a synchrotron is fired towards a crystal of known atomic spacing of planes. Each wavelength of the beam is diffracted off the monochromator crystal at a different angle. By rotating the crystal in line with the specific angle of diffraction for a particular wavelength, light of such wavelength may be selected for use in an

experiment. **9** 2.7\AA

10 a A number of peaks appear in the graph to correspond with various Bragg maxima of $n = 1, n = 2, n = 3$, etc. **b** As the horizontal axis does not start from zero, we estimate that the first maxima ($n = 1$) occurs at a 2θ angle of 16° ; so $\theta = 7^\circ$. So 2.9×10^{-10} m. Also from the graph, the second maxima seems to lie at a 2θ angle of approximately 27° . This means $\theta = 13.5^\circ$. So 3.0×10^{-10} m. The third maxima seems to lie at a 2θ angle of approximately 31° . This means that $\theta = 16.5^\circ$. So 3.7×10^{-10} m. In averaging these three figures, we can estimate the spacing between atomic planes to be 3.2×10^{-10} m, or 3.2\AA .

13.4 Scattering and beyond

1 C **2 C** **3** Blue light has a higher frequency than red light, because it has photons of greater energy.

4 a 3.3×10^{-19} J **b** 6.4×10^7 m s⁻¹ **5** X-ray photons undergoing a Compton collision are involved in an inelastic interaction. The scattered photon then emerges with less energy than before. Because energy, $E = hf$, then the photon emerges with lower frequency and a correspondingly longer wavelength than unscattered photons that emerge unchanged.

6 The diffuse scattering of partially ordered or amorphous samples produces a low intensity of scattered X-rays. These can provide information about variation in a crystal lattice structure and data regarding the movement of molecules within a crystal.

7 a Powder diffraction **b** Small-angle scattering **c** Magnetic Compton scattering **d** Diffuse scattering **e** Extended X-ray absorption fine structure.

8 a 5.2×10^{-10} m **b** 3.8×10^{-16} J **c** 1.9×10^{-16} J **d** 1.9×10^{-16} J **e** 2.0×10^7 m s⁻¹

9 a The outgoing X-ray photon had longer wavelength, and so lower energy than the incoming photon used in the collision. **b** A photoelectron would have been emitted from the graphite sample which possessed kinetic energy equivalent to the energy loss of the incoming photon.

10 a 1.28×10^{-26} kg m s⁻¹ **b** By realising that photons possess momentum, Compton observed the fact that, in the case of Compton collisions, individual photons interact with electrons, transferring energy in the same way that particles undergo collisions, and obeying the same laws.

Chapter review

1 A 2 C 3 B 4 C 5 D 6 D 7 B 8 C 9 A 10 A 11 Electrons should be drawn to follow a circular path with magnetic force acting towards the centre.

12 5.8×10^{-2} m **13** For $n = 1$, $\theta = 13.8^\circ$. For $n = 2$, $\theta =$

28.4°. For $n = 3$, $\theta = 45.6^\circ$. For $n = 4$, $\theta = 72.2^\circ$. If $n = 5$, $\sin\theta > 1$; therefore, there are no more diffracted beams.

14 The map reveals the electron density of the sample due to scattering by outer electrons. Because hydrogen has only one electron per atom, the effect is minimal and these atoms do not show up on the map.

15 The region of greatest electron density is represented by the greatest bunching of contours. The atom with the greatest number of electrons is chlorine, with atomic number 17 as compared to 8 for oxygen and 4 for carbon.

16 To observe a diffracted beam, Bragg's law must be satisfied, i.e. $2d\sin\theta = n\lambda$. For $n = 1$, $\lambda = 4.01 \text{ \AA}$. This is not in the range of incident wavelengths. For $n = 2$, $\lambda = 2.01 \text{ \AA}$. For $n = 3$, $\lambda = 1.34 \text{ \AA}$. For $n = 4$, $\lambda = 1.00 \text{ \AA}$. For $n = 5$, $\lambda = 0.80 \text{ \AA}$. For $n = 6$, $\lambda = 0.67 \text{ \AA}$. This is not in the incident range. Therefore, you would expect to see diffracted beams for the fourth and fifth maxima.

17 $58 \times 10^6 \text{ m s}^{-1}$ **18** $7.0 \times 10^{-11} \text{ m}$

19 We would expect the X-ray photon to emerge with a longer wavelength because it has lost energy in the Compton collision to an ejected photoelectron.

20 The data shown give us information about the relative intensity of diffracted beams for various angles of 2θ . Using Bragg's law: $2d\sin\theta = n\lambda$ through a knowledge of the number of the maxima, n , and the wavelength of X-ray photons used in the experiment, the distance, d , between the layers of atoms in the sample may be calculated. Comparisons of the peak locations and sizes can also be referenced by using data from the ICDD.

Chapter 14

14.1 Incoherent light sources

1 The light emitted by the first globe is in random phase with respect to the second. For interference to be observed, the light sources must be in phase, as the original light from the single source split by the two slits. **2** D

3 The emitted light is related to the temperature of the radiator according to Wien's law: $\lambda_{\text{max}} = \alpha/T$. This means that the higher the temperature, the shorter the wavelength of the emitted radiation and the closer to the blue part of the spectrum the visible wavelengths of EMR will be. **4** B **5** 1.82 eV

6 1088 nm. As visible wavelengths range from 390 nm to 780 nm, silicon could be used, since visible light cannot be absorbed by the valence electrons.

7 1.24 eV or less **8** A **9** D

14.2 Coherent light sources: Lasers

1 $3.72 \times 10^5 \text{ km}$ **2 a** $4.74 \times 10^{14} \text{ Hz}$ **b** $3.14 \times 10^{-19} \text{ J}$
c 7.96×10^{15} **d** $3.0 \times 10^4 \text{ km}$

3 The process in which an atom that has been excited to a metastable state is struck by a photon with an amount of energy equal to the difference in energy level for the excited atom. This causes two identical photons in phase with each other to be emitted. Each of these photons can then interact with other excited atoms, releasing further identical and in phase photons. Coherent laser light is thus produced.

4 Mirrors placed at the end of the laser tube reflect the photons back and forth through the lasing medium. This induces more stimulated emissions of photons to occur. One mirror is only partially silvered so a beam of photons emanates from this end.

5 a 0.11 eV, $1.76 \times 10^{-20} \text{ J}$ **b** $1.13 \times 10^4 \text{ nm}$

c No, it produces infrared photons.

6 a No, UV light is not visible **b** electrons **c** 6.4 eV

7 The photons that are emitted from a light globe have a wide range of frequencies, while lasers produce photons with a single frequency. Photons from light globes are spreading out in all directions, while laser light is a very narrow beam. Light globes produce incoherent light, whereas laser light is coherent: the photons are in phase.

8 a 20.7 eV **b** $n = 3$ **c** $6.21 \times 10^{-7} \text{ m}$

14.3 Optical fibres

1 High bandwidth, low signal attenuation, size, electrical isolation, security. **2** A **3** 50 μm

4 a Mode 2 **b** Mode 1 **5** 2, 3, 1

6 Material dispersion. The LED has a spectral spread of frequencies that will be transmitted about the mean frequency being used. Each of these frequencies has a different velocity due to a slightly different refractive index in the glass core. As a result, the varying modes will arrive at the exit end of the fibre at different times, increasing the degree of pulse spreading.

7 0.2284 **8** 13.20°

9 The numerical aperture remains the same, but the acceptance angle decreases.

10 The curve in Figure 14.30 illustrates how wavelength affects signal attenuation in fibres. The two peaks in attenuation correspond to resonances with O-H bonds because of water contamination. Generally, the level of attenuation decreases with increasing wavelength, up until about 1.5 μm . The useful minima for signal losses are found around wavelengths of 1300 and 1550 nm. There is a rapid increase in attenuation in the infrared region due to absorption by Si-O and Ge-O bonds.

11 1.49 mW

14.4 Applications of optical fibres

1 96% **2** Single-mode fibres have low dispersion, even over long distances, with many advantages over copper cabling. It is currently too expensive to use optical fibres over short distance due to the cost of coupling devices and associated costs in setting up the system.

3 Far greater bandwidth **4** 0.24

5 Although far superior in bandwidth capabilities and in terms of pulse dispersion, single-mode fibre is more expensive to install and connect to a network than multimode fibre. Over a short distance pulse spreading is limited, so single-mode fibre is an effective choice.

6 A movement or pressure variation causes a minute movement in the mirror from one position to another, as shown in the diagram. As the distance between the mirror and the fibre end-faces increases, the position of the cone of reflected light from the feed fibre changes and the amount of light gathered by the return fibre increases. A comparison of the signals from the return fibre (in the form of signal processing) can equate the differences with the physical movement or phenomenon that produced the variation.

7 In the case of an extrinsic sensor, the physical parameter that is to be measured interacts with the light from the feed fibre outside the fibre itself. An example is the moving-reflector sensor, which can be used to detect minute variations in pressure. In the case of an intrinsic sensor, the physical parameter interacts with the light within the fibre itself. An example is the oil-in-water detector used in oil tankers or the sensor for physical change making use of microbending.

8 66.5°

9 If oil has leaked into the water, oil droplets will attach themselves to the exposed glass core. As light travels through the sensor modulation zone, instead of being totally internally reflected along the core–water boundary some of it will be refracted into the oil. This occurs because the refractive index of the oil (1.50) is greater than that of the core (1.45). The sensor will detect a loss of signal intensity and translate this into information about the concentration of oil in the water.

10 The endoscope utilises two non-ordered and one ordered light guide. The fibres in the ordered light guide have been aligned into an ordered array. These are able to transmit the image of the object being studied with the endoscope. The fibres in the non-ordered guides are not aligned and will just conduct light. This is necessary to illuminate the object being viewed.

Chapter review

1 a B **b** A **c** C **d** D **2 a** A **b** B **c** C **3** C **4** A
5 C **6** D **7** E **8 a** 3.5×10^{17} **b** $6 \times 10^{-3} \text{ m} = 6 \text{ mm}$
c 5.8×10^{16}

9 The energy released by the atom was obtained from bombarding electrons or photons, the ‘pumping’ energy, which raised its energy level. The stimulating photon gives no energy to the atom.

10 The atoms release photons in all directions. The mirrors ensure that the vast majority of photons are released in the required directions by reflecting only the ones travelling in the appropriate direction. These then cause the chain reaction to occur, building up the number reflected back and forth. The mirror at one end is not fully reflecting and allows a portion of the light to escape.

11 Because laser light is perfectly parallel, it can be focused to a tiny point, the size of which is only limited by diffraction effects. Light from any other source can only be focused to a small image of the source. In addition, laser light is monochromatic so there are no dispersion problems in the lens.

12 Light is an electromagnetic wave with a frequency of the order of 10^{14} Hz, while the signals in wires are limited to frequencies of less than around 10^9 Hz (usually much less). Ultimately a wave cannot carry more information each second than its frequency. However, many practical factors limit the actual carrying capacity or ‘bandwidth’ to well below the theoretical maximum.

13 Any moisture or grease on the surface would change the relative refractive index, and hence the critical angle, and so lead to loss of light.

14 The cladding must have a lower refractive index, so that total internal reflection occurs. However, if the cladding has a much lower refractive index, light will reflect at larger angles and there will be many paths available, which will result in the spreading out of the light pulses.

15 The devices that convert the optical signals to electrical signals are complex and expensive. It is easy to connect an electrical cable to systems in the home, but not an optical fibre. For this reason, the optical fibre is terminated in an exchange where the messages and signals are separated and sent down individual copper cables to the appropriate end users.

16 a 21.0° **b** 32.5° **c** 32.5° **d** 7.1% longer **e** 0.538

17 A step-index fibre has a uniform refractive index in the core with an abrupt decrease at the cladding. Graded-index fibre has a higher refractive index in the centre of the core, decreasing toward the outside

cladding. In step-index fibres the light all travels at the same speed, and so modal dispersion occurs. In graded-index fibres the light in the outer part travels faster than the light in the centre, and this compensates for the longer path, thus producing less modal dispersion.

18 Over short distances dispersion (of both types) is less important, because the pulses do not have time to spread out much. Therefore, larger cables are used, because it is easier to couple them and to get more light into them, thus making them cheaper to use. Over long distances dispersion becomes greater due to the longer time the pulses are travelling, and so single-mode fibre, which must have a core diameter only a little larger than the wavelength of the light, is used to eliminate modal dispersion.

19 a 0.300 **b** The acceptance angle depends on the medium surrounding the fibre, while the NA is dependent only on the fibre itself. **c (i)** 17.5° **(ii)** 13.1°

20 The main limitations on the capacity of an optical-fibre system are not so much the frequency of the light but the dispersion (modal and material) and attenuation (loss of signal). UV light is absorbed and scattered by glass much more than IR light, leading to greater attenuation. Also, the smaller wavelength of UV light would mean that the diameter of the fibres would have to be smaller to reduce modal dispersion.

21 a 50 MHz **b** 10 m

22 The speed of light depends on the refractive index ($v = c/n$). In this glass, the time for a violet pulse to travel 1 km is given by $t = d/v = 10^3 / (3 \times 10^8 / 1.532) = 5.11 \times 10^{-6} \text{ s} = 5.11 \mu\text{s}$. For red light this figure is $5.04 \mu\text{s}$. Thus, over this distance the pulse would have spread out over a time of $0.07 \mu\text{s}$ and so pulses would need to be separated by at least this time. This corresponds to an upper limit frequency of 14 MHz. This is an example of 'material dispersion'. **23** B–E **24** B

25 At smaller wavelengths Rayleigh scattering occurs, which results in considerable attenuation. At longer wavelengths the silica absorbs the infrared light. At 1400 nm there is a peak where water molecules, which unavoidably contaminate the glass, absorb IR radiation.

26 LEDs last long and are cheap, but they emit light with a small spread of wavelengths which leads to material dispersion. Also, because they emit light over a relatively large range of angles, it is more difficult to couple their light into the fibre. While lasers are more expensive and don't last as long, they emit a very sharp spectral line and their beam is easier to couple into the fibre.

27 The optical repeater does not require the signal to be removed from the optic system, converted to an electric signal, amplified and cleaned up and then converted

back into an optical signal. All these processes are expensive and result in some losses. However, the optical repeater amplifies the whole signal including distortions and spreading, while the regenerator takes the incoming signal, removes the spreading and losses, and produces a new clean, sharp pulse. **28** C

29 Answers could include: loss of light through bending of the fibre, loss of light through temperature-dependent refractive indices, loss of light through changes in the refractive index of the material surrounding the core, loss of light by motion of a mirror at the end of a fibre, a change in the frequency or phase of the light, a change in the modal pattern of the light. There are many other possibilities as well.

30 Light is sent down the unordered bundle(s) of fibres to illuminate the object with either visible or perhaps UV light. A lens focuses the image onto the face of the ordered bundle and this transmits the image to another lens or optical system at the viewing end.

Chapter 15

15.1 The nature of sound

1 C **2** C, E **3** A **4** B

5 In a transverse wave, the particles vibrate at right angles to the direction of wave travel. In a longitudinal wave, the vibration is in the same direction as the direction of propagation. **6** 0.60 m

7 a The microphone converts sound energy (kinetic energy of vibrating air molecules) into electrical energy.

b $t = 0.5 \text{ ms}, 1.5 \text{ ms}, 2.5 \text{ ms}, 3.5 \text{ ms}, 4.5 \text{ ms}, 5.5 \text{ ms}$

8 a P, R and T are points of maximum positive pressure variation. **b** Q and S are points of maximum negative pressure variation. **c** This is a point of zero pressure variation. **d** This is a point of zero pressure variation.

9 a $t = 0.5 \text{ ms}, 2.5 \text{ ms}, 4.5 \text{ ms}$

b $t = 1.5 \text{ ms}, 3.5 \text{ ms}, 5.5 \text{ ms}$

15.2 The wave equation

1 C **2** There is a greater force of attraction between the molecules in water than in air and when a water molecule is displaced, it returns to the mean position faster. This means that the wave will travel faster.

3 C

4 D. It could also be argued that since $f = v/\lambda$ and $E \propto \rho A^2 f^2 v$, B and C will also be correct and, as far as A depends on C, A is as well. **5** C **6** 300 m s^{-1} .

7 0.059 s **8** 340 m s^{-1}

9 a 1.5 m **b** 1500 m or 1.5 km **c** The medium in which the sound is travelling does not effect the frequency; hence, ratio is 1 : 1.

10 a 0.75 m **b** Speed of signal remains constant since

the medium remains unchanged. The time is therefore the same.

15.3 Diffraction of sound

- 1** 0.34 m or 34 cm **2** Since $\lambda \approx w$, appreciable diffraction will occur, resulting in significant sound energy arriving at points P and Q.
3 0.085 m or 8.5 cm
4 λ/w is much less than for $f = 1000$ Hz, so much less diffraction will occur and less wave energy will arrive at these points.
5 Higher frequency sound is diffracted less, so more of the wave energy will travel directly to Q. **6** B. **7** D.
8 High-frequency sound undergoes less diffraction as λ is smaller, so the sonar waves will tend to travel directly to and from the object in a beam with less spread than audible frequencies. **9** A
10 The violin—since this instrument produces sound of a higher frequency, diffracting less as it passes through the door.

15.4 Amplitude, intensity and the decibel scale

- 1** D **2** While the energy from the sound wave would be very similar, the response of two different individuals' ears to the same sound could be quite different, due to difference in perception of a particular frequency or hearing loss. **3** E
4 a 0.23 m or 23 cm **b** 23 cm **5 a** $4.4 \times 10^{-3} \text{ W m}^{-2}$
b $1.6 \times 10^{-3} \text{ W m}^{-2}$
6 a $8.8 \times 10^{-3} \text{ W m}^{-2}$ **b** $3.2 \times 10^{-3} \text{ W m}^{-2}$ **7** 10 J
8 a 109 dB **b** 99 dB **c** 95 dB
9 a $1.0 \times 10^{-2} \text{ W m}^{-2}$ **b** $1.1 \times 10^{-3} \text{ W m}^{-2}$ **c** 90 dB
10 $8.0 \times 10^{-9} \text{ W m}^{-2}$ **11 a** $1.0 \times 10^{-5} \text{ W m}^{-2}$ **b** 70 dB

15.5 Frequency, perceived loudness and the phon

- 1** Frequency is determined by the source of the sound and is independent of the listener. Pitch is how the frequency of the sound is perceived by a particular listener.
2 D **3** C
4 A (B is also correct but A is the better answer in the context of the question.)
5 a 200 Hz **b** 100 Hz **c** 50 Hz **6** 90 dB
7 400–2000 Hz **8** C
9 a $1.26 \times 10^{-7} \text{ W}$ **b** 25 dB **10** 40 dB

15.6 Making sound: strings and air columns

- 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 will then increase. If sufficient energy is supplied, the amplification from resonance will cause the glass to 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** At a point one-quarter of its length from either end **c** At a point one-sixth of its length from either end
7 a 300 Hz **b** 600 Hz **c** 900 Hz
8 Resonance in air columns of a particular length is due to reflection of waves arriving at the ends of the column. The reflected waves are superimposed on the existing waves to produce a standing wave pattern. This results in resonance.
9 a 0.90 m **b** 0.45 m **c** 1.1 kHz
10 a 110 Hz **b** 330 Hz **c** 550 Hz and 770 Hz

15.7 Recording and reproducing sound: the first and last links

- 1** A moving coil loudspeaker typically uses a flexible cone to produce sounds, while a moving coil microphone uses a shielded diaphragm. The flat surface of the diaphragm provides a more reliable surface for converting sound waves to an electrical current.
2 The larger the loudspeaker, the larger the cone and the more difficult it would be to make it vibrate with incident sound waves. Also, the 3D nature of the cone would make accurate reproduction of incident sound waves difficult. Thus a small, relatively flat cone—typical of small mid-range or tweeter speakers—would be the most suitable.
3 The small, flat cone of a tweeter could not vibrate at the low speeds necessary to reproduce low frequencies reliably.
4 Dropping the cone onto a bench would cause the coil of wire glued to the cone to move within the magnetic field of the magnet. An induced voltage would result and would be seen as a jump in the trace of an oscilloscope.
5 A cylindrical magnet produces a strong radial field in the gap between north and south poles. The coil of the speaker vibrates within this field. The stronger the magnetic field, the higher—and therefore more distinct—the induced current/voltage.
6 Their frequency range is too limited. (Carbon

microphones cannot respond to frequencies above 4 kHz.)

7 The combined frequency responses of the different speakers need to cover the full range of frequencies audible to humans, while accurately producing particular small groups of frequencies.

8 The electrodynamic, or moving coil, microphone provides the broad frequency response required for recording music.

9 A small diaphragm is attached to a coil of wire. When sound energy is incident on the diaphragm, it moves up and down, moving the coil between the poles of a permanent magnet. This generates an induced voltage of varying frequency and amplitude in the coil, directly related to the variation in sound energy incident on the microphone's diaphragm.

10 As both microphones and loudspeakers convert one form of energy to another, both could be accurately described as transducers.

Chapter review

1 C 2 B 3 C 4 A 5 C 6 A 7 C 8 A 9 D

10 C 11 C, D 12 100 Hz 13 300 Hz

14 2.0×10^{-3} W 15 4.0×10^{-7} Wm⁻² 16 1.2 J

17 Intensity of sound is a measurable quantity (units W m⁻²), while the loudness of a sound depends on how a particular individual's ears respond to that sound and can vary with frequency for each individual.

18 An increase of 3.0 dB in sound level represents a doubling in sound intensity. Prolonged exposure to this increased level could cause problems.

19 Most participants report the greatest difference in perceived loudness at frequencies below 500 Hz and above 5.0 kHz.

20 A tweeter best produces frequencies between 4 and 20 kHz. It would distort sounds, and test results, for frequencies significantly below 4 kHz. For best results, a broad range speaker with a flat frequency response curve for frequencies between 20 Hz and 20 kHz (the range of normal human hearing) would be required.

21 Baffling reduces, but rarely eliminates, reflections of sound waves that are created by the movement of the back surface of a loudspeaker's cone. If not reduced, these sound waves would be superimposed on the sound waves from the front of the speaker cone to cancel out, or reduce, some frequencies.

22 Electret-condensor. Mobile phones need to reliably reproduce the frequency range of the human voice—generally below 1 kHz. The microphone also needs to be compact and durable. The small capsule, low cost and durability of an electret-condensor microphone would be a good option. Its poor response to high frequencies would not be a drawback for this application.

absolute quantities Physical quantities that remain constant within any frame of reference.

absorption spectrum When a beam of white light (a continuous spectrum) is passed through a sample of a gas, particular frequencies of light are missing, indicating that some frequencies have been removed by the gas. The absorption spectrum is unique to each type of atom and is due to the absorption of energy as electrons move to higher energy levels within the atom.

acceptance angle The maximum angle at which a ray can enter an optical fibre and then propagate by total internal reflection.

action/reaction forces Force pairs used in Newton's third law to describe the forces that bodies exert on each other during collisions and interactions. These forces are always equal and opposite to each other.

aether The supposed medium through which light waves (see *electromagnetic radiation*) were supposed to travel just as, for example, sound waves travel through air. Einstein decided it was an unnecessary concept and that light needs no physical medium in which to travel.

air column In sound, usually a vibrating column of air contained within the pipe of a musical instrument.

amplifier clipping Non-linear amplification that results in signal distortion, in which the peaks of the output waveform are limited (usually to the positive and/or negative power supply voltage).

amplitude The maximum absolute value of a periodically varying quantity such as a transverse or longitudinal wave.

anode The electrode maintained at a positive potential. Attracts electrons in a discharge tube.

antinode The point along a standing wave where the wave has maximum amplitude. This may be in terms of displacement or pressure variation depending upon the quantity being investigated.

aphelion Point in an elliptical orbit that is furthest from the Sun.

apparent weightlessness The state of a body that is falling freely in a gravitational field. The body will move with an acceleration that is equal to the gravitational field strength, g , at that location.

attenuation (photonics) Loss of signal strength—in the case of optical fibres, optical power—along the length of the cable.

beamline Pathway in which a photon beam generated in a synchrotron travels from storage ring to an experiment room, or endstation.

BJT transistor A bipolar junction transistor made by connecting p-type and n-type semiconductor materials to form two interacting pn junctions. The base current is amplified to give a larger collector current via transistor action.

booster ring A booster synchrotron which operates to accelerate electrons to sufficiently high energy before being injected into the storage ring of a synchrotron.

branch current Current flowing in a particular branch of an electronic circuit.

brittle A material that only displays elastic behaviour and fails once the elastic limit is exceeded.

cantilever A beam that extends beyond its support structure.

capacitance Measure of the capacity of two separated conductors to store charge; defined as the ratio of the magnitude of the charge on either plate to the potential difference between them.

capacitor An electronic device designed to have a specific capacitance.

cathode ray oscilloscope (CRO) An electronic measuring instrument that can display AC and other periodically varying voltages.

cathode ray tube An evacuated glass tube within which electrons are accelerated from a source and strike a fluorescent screen.

cathode The electrode maintained at a negative potential. Emits electrons in a discharge tube.

centre of mass The point at which the entire mass of an extended body can be considered to be situated for the purpose of analysing its motion.

centripetal acceleration An object moving with a constant speed in a circular path has a centripetal acceleration towards the centre of the circle.

centripetal force The force that causes an object to travel in a circular path; can include gravity, tension and friction.

circuit ground or **earth point** Zero electrical potential reference point.

coherent Light that is in phase. The photons are synchronised with each other. Laser light is coherent light.

commutator The rotating cylindrical copper segments in an electric motor that carry the current from the brushes to the coils in the armature, which are in the best position for maximum torque.

composite material A material such as reinforced concrete and consisting of two or more materials, employing the useful properties of each.

compression or **compressive force** Force acting in a material that has been squeezed or squashed.

Compton scattering The result of an inelastic collision between an X-ray photon and an electron, in which some energy is transferred to the electron and the X-ray photon continues with a longer wavelength.

conservation of energy A fundamental principle of nature, which states that energy can be transformed from one form to another, but that the total amount of energy remains unaltered, i.e. energy cannot be created or destroyed.

conservation of momentum During any collision, the total momentum of the bodies involved in the collision remains unchanged.

control room The shielded area in which scientists can monitor and control their experiment using synchrotron light from a safe distance.

corpuscles Isaac Newton interpreted rays of light as streams of tiny particles or 'corpuscles' travelling at high speed.

current or **conventional current** The transfer of positive charge with respect to time. Electron flow is in the opposite direction to electron flow.

cyclotron An early type of circular accelerator, invented in 1929.

dark current Small current (due to the thermally generated charge carriers) that flows when a photosensitive semiconductor has a potential difference across it. It is always present even when the semiconductor is in total darkness. See also thermal leakage current.

de Broglie wavelength Also known as the 'matter wavelength'. This term refers to the wave/particle duality of matter whereby very fast moving particles display wave-like properties and can be allocated a de Broglie wavelength through observation of their diffraction patterns, for example.

depletion region The region around a pn junction that has virtually no mobile charge carriers.

diffraction When a wavefront meets an aperture or obstacle of a similar dimension to the wavelength, the waves spread around the aperture or obstacle, effectively bending the direction of travel of a section of the wave.

diffuse scattering The pattern produced by weaker Bragg peaks as a result of X-ray diffraction from a sample that is lacking a well-ordered crystalline structure.

diode A non-ohmic electronic device that conducts well in one direction but poorly in the other.

dipole field The magnetic field around a magnet appears to come from the two (di-) poles at the ends of a magnet. The field around a current has no poles and so is a 'non-dipole' field.

doping The process of increasing a semiconductor's conductivity significantly by the addition of impurity atoms in the semiconductor lattice.

ductile A material that can withstand stresses greater than the elastic limit and will undergo significant plastic behaviour before failing.

elastic behaviour When a material is subjected to stresses lower than the elastic limit. Once the applied stress is removed, the material returns to its original dimensions.

elastic limit The point at which the F - x graph or stress-strain graph for an elastic material starts to behave in a non-linear manner. The material will experience permanent deformation if the elastic limit is exceeded.

elastic potential energy Stored energy in a stretched or compressed material. Also known as strain energy and measured in joules (J).

electroluminescence The mechanism that allows the *light-emitting diode* (LED) to generate optical radiation.

electromagnet A magnet with a field produced by an electric current.

electromagnetic induction The creation of an electric current (or an EMF) in a loop of wire as the result of a changing magnetic flux through the loop.

electromagnetic radiation Energy emitted in continuous waves with two transverse, mutually perpendicular components: a varying magnetic field and a varying electric field.

electron diffusion When mobile electrons move from a region of high electron concentration into a region of low electron concentration.

electron gun The source of electrons in a cathode ray tube or particle accelerator, consisting of a hot filament from which electrons are emitted and focused into a beam.

emission spectrum The particular set of frequencies of light, unique to an element, emitted by an excited atom as the excited electrons return to lower energy levels.

energy levels of atoms The orbital levels in which the electrons orbiting the nucleus of an atom can remain stable.

experiment room or **endstation** The room at the end of a synchrotron beamline in which a specific experiment occurs.

extrinsic sensors Fibre-optic sensor modulation process is, as the name suggests, external to the fibre itself. Light is taken out of the feed fibre, modified in some manner, and then coupled back into the return fibre.

ferromagnetic A material in which tiny 'domains' of atoms combine to form a strong magnetic field. Iron is the most common ferromagnetic material.

forward bias voltage Voltage across a pn junction that greatly increases its conductivity.

frame of reference A system of coordinates (e.g. x - y - z), usually attached to some physical system, from which to measure the relative motion of another system. An inertial frame of reference is one which is not accelerating.

free-fall A motion whereby gravity is the only force acting on a body.

frequency The number of waves passing a given point in one second. Measured in Hertz (Hz), or cycles per second.

frequency response A measure of what frequencies can be reproduced and how accurately they are reproduced.

gamma See *Lorentz factor*

GPS [geographic positioning system] System of satellites, originally developed for the US military, for finding position anywhere on Earth.

gravitation Force of attraction between two objects. Also known as gravity.

gravitational field The region around an object where other objects will experience a gravitational force.

gravitational potential energy Energy that can be considered to be 'stored' in a body due to its position in a gravitational field. A scalar quantity that is measured in joules (J).

heliocentric Sun-centred.

Hooke's law Elastic materials for which there is a direct relationship between the force acting on them and the extension or compression that they undergo are said to obey Hooke's law.

ideal battery A battery that maintains its terminal voltage regardless of how much current is drawn from it.

impulse The product of an applied force and the time interval for which it is applied. Impulse is equal to the change in momentum of the body. Impulse is a vector and is measured in newton seconds (N s).

in phase Two or more points in waves of the same frequency that at a given point in time have the same displacement and velocity.

incandescent The kind of light that emanates from standard light bulbs.

inclined plane Sloping surface or ramp.

incoherent (optics) Electromagnetic radiant energy not all of the same phase, and possibly also consisting of various wavelengths.

insertion device Can be either a wiggler or an undulator, and consists of an array of magnets, which when placed in a straight section of a storage ring greatly increases the brightness of synchrotron light produced.

integrated circuit A chip of semiconductor on which many elements (transistors, resistors, capacitors) are combined to form a useful circuit.

intensity The energy incident per second on 1 m² of surface area; measured in watts per square metre (W m⁻²).

interference pattern The pattern due to light (or fast moving particles) from two or more sources arriving at the one location.

interference The effect on the amplitude or intensity at a given location due to the superposition of two or more waves incident upon the location.

interferometer A device which combines two beams of light in such a way they *interfere* and produce an *interference pattern* of light and dark bands, enabling the measurement of very small distances in terms of the wavelength of light.

internal forces Forces exerted on each other by bodies involved in a collision, e.g. action–reaction forces.

Internet Distributed digital communications system that uses fibre optic cable (and other technologies) to transmit information in an asynchronous manner.

intrinsic sensors Fibre-optic sensor modulation process is, as the name suggests, internal to the fibre itself. Light is modified within the fibre feed.

isolated system A group of bodies interacting with no influence from any external forces that are changing the energy of the bodies.

Kepler's laws Three rules devised by Johannes Kepler that describe the orbital motion of planets.

laser (Light amplification by stimulated emission of radiation) A light source that has a very narrow emission spectrum (i.e. it emits a very pure colour), where all the wave components of the light are in phase.

laser diode (LD) Similar in concept to an LED but emits laser light. Usually has a very narrow spectral range of around a few nanometres or less.

length contraction An object moving fast in our frame object will appear shortened by a factor equal to the *Lorentz factor*. Dimensions at right angles to the direction of motion are not affected.

light-dependent resistor (LDR) Semiconductor device whose resistance varies significantly as the illuminating light level changes.

light-emitting diode (LED) A pn junction diode made of Ga, As or similar semiconductor material. When forward biased, the diode emits light. Usually has a narrow spectral range of a few tens of nanometres (i.e. emits one single colour).

line current Total current flowing from the battery into a circuit.

linear accelerator (linac) Type of particle accelerator in which particles are accelerated in a straight line.

longitudinal wave A wave in which the vibration is moving in the same direction as that in which the wave is travelling.

Lorentz factor Also known as gamma (γ). This is the factor by which relativistic effects change normal measurements (for example time, length and mass) due to the relative speed between *frames of references*. It is extremely close to 1 for normal speeds and only reaches 1.01 at almost 15% of c .

loudness The overall quantity of sound, as perceived by a listener. Also referred to as volume or strength.

magnetic declination The difference between true north and the direction of the Earth's magnetic field at any point. In south-eastern Australia it is around 10° east.

magnetic field A region of space around a magnet or electric current in which another magnet or electric current will experience a force. Field lines point away from north and towards south magnetic poles.

magnetic flux The product of the strength of the magnetic field and the perpendicular area over which it is spread. A measure of the total 'amount' of magnetic field.

magnetic pole The ends of a magnet that the magnetic field lines appear to point to (south) and away from (north). There are no magnetic poles around electric currents.

magnitude The size of a quantity without regard for its direction.

material dispersion Because the light used in any fibre-optic communication system has a range of wavelength components, these will travel at slightly differing speeds. Hence, material dispersion occurs and any narrow pulse is spread.

matter wavelength Also known as the 'de Broglie wavelength'. Very fast moving particles display wave-like properties and can be allocated a de Broglie wavelength through observation of their diffraction patterns, for example.

matter waves Extremely fast moving particles that display wave-like properties, e.g. fast moving electrons will diffract as they pass between the atoms in a crystal structure.

modal dispersion Any sharp pulse of light entering an optical fibre will become more spread out as it travels further. This effect limits how close the pulses can be before they overlap and become indistinguishable from each other. The spreading out—or loss of definition—of the pulses is referred to as modal dispersion.

momentum of a photon An interpretation of light describes the momentum of a photon as dependent upon the frequency (and therefore wavelength) of the photon. If

- photons can be described as 'having momentum', this suggests a particle-like nature of light.
- momentum** The product of the mass of an object and its velocity—a vector quantity, measured in kg m s^{-1} .
- monochromatic light** Light of a single frequency (and wavelength).
- monochromator** A device, commonly a crystal, used to diffract and select out particular wavelengths of synchrotron light.
- multimeter** An instrument that can measure voltage difference, current and resistance.
- multimode fibre** A fibre optic carrying more than one mode of propagation.
- natural satellite** A body such as the Moon or a planet that is in orbit around another body.
- negative charge flow** Flow of mobile negative charge or electrons around a circuit. The flow is regulated by the electrical potential difference and the electrical resistance to the flow.
- node** A node is a point along a standing wave where the wave has minimal amplitude. This may be in terms of displacement or pressure variation depending upon the quantity being investigated.
- non-ohmic device** An electrical device whose resistance changes as the applied voltage changes.
- n-type semiconductor** Doped semiconductor that easily generates free negatively charged electrons in the lattice.
- nuclear model** A model of the atom in which the majority of the mass of the atom is concentrated in the central nucleus, with the electrons orbiting the nucleus.
- numerical aperture** A measure of how readily it will capture light. It has a value between zero and one. $NA = n \sin \theta_1$.
- ohmic device** An electrical device whose resistance is independent of the applied voltage.
- ohmmeter** An instrument that can measure the electrical resistance of a particular circuit element. The element must be electrically disconnected from the rest of the circuit.
- optical attenuation** Reduction in propagating optical power as a signal travels through a medium.
- optical fibre** Any filament or fibre made of dielectric materials, that is used to transmit laser or LED-generated signal
- optical spectrum** Part of the EM spectrum that ranges in wavelength from the non-visible UV through the visible light spectrum (violet to red) into the non-visible IR.
- optics room** Area in which synchrotron light is modified to make it suitable for use in a particular experiment in a beamline.
- particle accelerator** a device that uses electric fields to accelerate charged particles to very high speeds.
- perihelion** Point in an elliptical orbit that is closest to the Sun.
- period** The interval taken to complete one cycle of a regularly repeating phenomenon such as a sound wave.
- phon curve** A graph of the level at which a sound is heard with apparent equal loudness.
- photoconductive detector** A light detector (e.g. an LDR) whose conductance varies significantly as the light level illuminating it changes.
- photodiode or junction optical detector** A pn junction semiconductor device that can absorb light into its depletion region and generate a measurable separation of charge.
- photoelectric effect** The emission of electrons from the surface of a material when light of sufficiently high energy shines on it. Energy is transmitted from a photon of light to a surface electron in quantised amounts.
- photoelectron** An electron released from an atom due to the photoelectric effect.
- photon** A packet of electromagnetic energy released from an atom as an excited electron falls to a lower energy level.
- photonics** The science of using light to manipulate information and energy.
- photophone** System for communicating voice signals using a light beam over a free air path. Invented by Alexander Graham Bell in 1880.
- phototransistor** A transistor device that can absorb light into its base region (creating a base photocurrent). This photocurrent is amplified to give a larger collector current via transistor action.
- pitch** The property of sound that varies with variation in the frequency of vibration.
- pn junction** An electronic structure that arises when a p-type semiconductor region is adjacent to an n-type semiconductor region.
- positive charge flow** Flow of mobile positive charge around a circuit. The flow is regulated by the electrical potential difference and the electrical resistance to the flow.
- powder diffraction** A technique in which X-ray diffraction studies are carried out on a powdered or microcrystalline sample.
- power dissipated** The rate of change of electrical energy with time. Also equal to the potential difference multiplied by the current.
- power** The rate at which work is done—a scalar quantity measured in watts (W).
- probability distribution** A mathematical wave function that gives the probability of a particular event occurring in a region of interest.
- projectile** Object moving freely through the air without an engine or power source driving it.
- proper length** Length as measured by observers in the same frame of reference as the object, i.e. by observers at rest relative to the object.
- proper time** Time between events as measured by observers in the same frame of reference as the events, i.e. by observers who see the events occur at the same place in their frame.
- p-type semiconductor** Doped semiconductor that easily generates free positively charged holes in the lattice.
- quanta** Discrete quantities. Usually refers to a discrete quantity of energy gained or lost by an atom during an electron level transition.

Rayleigh scattering The elastic scattering of light or other electromagnetic radiation by particles much smaller than the wavelength of the light (named after Lord Rayleigh).

real battery Battery where the terminal voltage decreases with increasing current drawn from it. Can be modelled by having an ideal battery in series with a fixed internal resistance.

rectifier A device that converts an AC voltage and current into DC. A half-wave rectifier blocks the negative AC cycles while a full-wave rectifier reverses the negative cycles.

relativistic mass At relativistic speeds, momentum increases with the Lorentz factor (γ). Thus it appears that the mass of an object is increasing with this factor. This apparent mass increase is referred to as the relativistic mass ($m = \gamma m_0$, where m_0 is the rest mass).

relativity All motion can only be measured relative to another *frame of reference*. There is no way to measure an absolute zero of velocity in the Universe. Galileo's principle of relativity expressed the practical nature of this concept. Einstein's special theory of relativity assumed it was a fundamental principle of nature and made it one of his two fundamental postulates.

resonance The natural tendency of a device to vibrate at a specific frequency. Unwanted resonances in loudspeakers, for example, alter the sound by producing excessive response at some frequencies.

reverse bias voltage Voltage across a pn junction that greatly decreases its conductivity.

reverse potential The external source of potential difference in a circuit is reversed, i.e. the positive and negative terminals switch locations.

RF (radio-frequency) cavities Regions in which a large, oscillating electric field is created and is used within a synchrotron to accelerate electrons.

right-hand rules 1. The right-hand grip rule tells us the direction of the magnetic field (curled fingers) around a current (thumb). 2. The right-hand force rule tells us the force (palm) on a current (thumb) in a magnetic field (straight fingers).

ripple voltage A small varying voltage on top of a DC voltage. It usually needs to be reduced by use of a *capacitor* or a *voltage regulator*.

RMS voltage and current Root mean square voltage/current of an AC supply is the equivalent DC voltage/current that would deliver the same amount of power. The RMS value is equal to the peak value of the AC quantity divided by $\sqrt{2}$.

rotational equilibrium When the sum of the torques acting on a structure is equal to zero.

satellite Object in a stable orbit around a central body.

shear If a sideways force acts across the top of an object while an opposing sideways force acts across the bottom, the material is under shear stress.

simultaneity The idea that if two events appear to occur at the same time to one observer, then they will also be simultaneous to any other observer. Einstein showed that for observers in relative motion, this will not be exactly true.

single-mode fibre An optical fibre that supports only one mode of light propagation above the cut-off wavelength.

skin effect How far electrical signals can penetrate into a conductor. Penetration depth decreases as the frequency of the signal increases.

small angle scattering The name given to a range of techniques in which electromagnetic radiation is elastically scattered off a sample and is examined to reveal information about the atomic structure of the sample material.

sound A form of mechanical energy transferred by the vibration of the molecules within the medium. Sound requires a medium in which to travel.

sound fidelity A measure of the ability of loudspeakers, microphones and other recording and reproduction devices to accurately reproduce an original sound.

sound intensity The sound intensity, I , (acoustic intensity) is defined as the sound power per unit area A . It is a measure of the amplitude of the sound wave.

sound levels Subjective measure of sound expressed in decibels as a comparison to the threshold of hearing for the average person.

spacetime A word which suggests that the concepts of space and time cannot be thought of separately in the relativistic world, but are linked in a four-dimensional reality.

spectrum A set of frequencies, usually of electromagnetic radiation.

spring constant (k) Gives a measure of the stiffness of an elastic material. The spring constant (or force constant) is given by the gradient of the F - x graph for the material.

standing wave When a physical object is subjected to a periodic driving force that matches a natural frequency of the object, it absorbs maximum energy from the driving force and it resonates. A standing wave is established in the object so that particular locations of maximum amplitude of vibration occur and locations of minimum disturbance also occur in set positions.

static equilibrium When a structure is in both translational and rotational equilibrium.

step-index fibre A multimode or single-mode optical fibre with a uniform refractive index throughout the core. The step is the shift between the core and the cladding, which has a lower refractive index.

stimulated emission Incident photon interacts with an excited electron and loses its energy by emission of a photon of identical phase direction and energy.

storage ring A circular tunnel within a synchrotron in which electrons can circulate for a period of time while maintaining a constant path radius and speed. It is within this region that synchrotron light is produced.

strain energy The work done in changing the length of a material. Strain energy in joules (J) can be found by multiplying the area under a stress-strain graph by the volume of the material under stress.

strain The amount of distortion (compression or extension) per unit length of a material under stress. Strain has no unit.

- strength** The maximum stress that a material can withstand before failing under tension or compression. Measured in pascals (Pa).
- stress** The applied force per unit cross-sectional area of a material when the material is under tension or compression. Stress is measured in newtons per square metre or pascals (Pa).
- strut** A rigid part of a structure that is under compression and acts to strengthen the structure.
- superposition** The interference of waves, or summing of amplitude's at the point of intersection.
- switch-on voltage** The voltage across the terminals of a diode beyond which the diode conducts strongly. For Si, $V_s \approx 0.7$ V; for Ge, $V_s \approx 0.3$ V.
- synchrotron** A type of particle accelerator in which electrons are accelerated by oscillating electric fields and bent by a series of bending magnets to follow a constant path radius. The accelerating particles produce a spectrum of electromagnetic radiation known as synchrotron light when travelling at sufficient speed and bending through an arc.
- synchrotron light** The broad spectrum of electromagnetic radiation ranging from infrared to hard X-rays as created in a synchrotron.
- temperature-dependent resistor (TDR)** An electronic device whose resistance varies as a function of temperature.
- tension or tensile force** Force acting in a stretched material.
- thermistor** Semiconductor device whose resistance varies significantly as the temperature changes. A thermistor is an example of a TDR.
- Thomson scattering** Elastic scattering of electromagnetic radiation by a charged particle.
- threshold frequency** The minimum frequency of electromagnetic radiation for which the photoelectric effect can occur for a given material; measured in hertz (Hz).
- tie** A rigid or flexible part of a structure that is under tension and acts to strengthen the structure.
- time dilation** When one observer watches events in a *frame of reference* that is moving (very fast) relative to him, time in that frame will appear to go more slowly. People in the moving frame do not experience any difference in the rate at which time passes. This is one of the strange consequences of Einstein's theory of special relativity.
- torque** When a force rotates an object, the product of the force and the perpendicular distance between the axis and the line of force is the torque (measured in N m).
- toughness** The ability of a material to absorb energy before it fails. Tough materials undergo significant plastic behaviour before failing.
- transducer** A device that converts electrical energy into non-electrical energy (light, mechanical, heat etc.) or vice versa.
- transformer** Two coils wound on an iron core so that the magnetic flux from one passes through the other. An AC current in one coil will induce an EMF, proportional to the ratio of turns, in the other and so can 'step up' or 'step down' the supply voltage.
- transistor** Three-terminal semiconductor device that can be used to amplify electronic signals.
- translational equilibrium** When the net force acting on a structure is zero, i.e. when the forces are balanced.
- transverse wave** A wave in which the vibration is moving perpendicular to the direction of wave propagation.
- true weightlessness** The state of a body in a region where the gravitational field strength is zero. An object would remain at rest or move with constant velocity in this region.
- turning effect** See torque.
- undulator** Type of insertion device consisting of a periodic array of many magnetic poles. Capable of enormously increasing the intensity of synchrotron light produced in a storage ring and greatly enhances specific wavelengths of the electromagnetic radiation produced.
- voltage amplification** The process of electronically magnifying a voltage signal.
- voltage gain** Ratio of output to input voltage. For linear amplification the ratio is constant.
- voltage regulator** An integrated circuit, the purpose of which is to take a varying DC voltage input and produce a constant definite voltage output.
- wave guide** A material medium that confines and guides a propagating electromagnetic wave, e.g. an optical fibre.
- wavelength** The distance between two identical points on neighboring waves.
- WDM** Wavelength division multiplexing, a type of multiplexing developed for use on optical fibres. WDM modulates each of several data streams onto a different part of the light spectrum.
- Wien's law** The peak wavelength of a wide-spectrum thermal radiator given by $\lambda_{\max} = \alpha/T$.
- wiggler** A type of insertion device consisting of alternating magnetic poles which greatly increases the intensity of synchrotron light generated when placed in a straight section of a storage ring.
- WLED or white-light-emitting diode** An LED that emits light over a broad spectral range and approximates a white light source.
- work** Done on a body when a force acts and changes the energy of the body. Work is a scalar and its unit is the newton metre or joule.
- work function** The energy required to remove an electron from its state of being bound to an atom; measured in joules or electronvolts.
- X-ray diffraction** X-rays diffract from a crystalline structure to produce a particular pattern. Analysis of this pattern reveals information about the spacing of atoms in the sample.
- Young's modulus (E)** The ratio of stress per unit strain for a material. This gives a measure of the stiffness or flexibility and is measured in pascals (Pa).
- zener diode** A silicon diode that is designed to have a relatively small, but stable, reverse-breakdown voltage. It is used to regulate a voltage in an electronic circuit.

- absorption spectra 448
- acceleration 4
 - centripetal 58–9
 - of falling objects 86
- acceleration–time graphs 7
- air resistance 16
 - effect of 27–8
- alternating current (AC) 283–4
 - generators 375
 - induction motors 383–4
 - and voltage 378
- alternators 375
- ammeters, analogue 353
- Ampere, André-Marie 333
- amplifiers
 - clipping in 139–40
 - transistor 140–1
 - voltage 137
- apparent weight 96
- apparent weightlessness 107–8
- arch structures 274–6
- atoms
 - Bohr’s model 449–50, 454–5
 - Rutherford’s model 446
- Australian synchrotron 464, 474–82
 - applications 348–9, 483
- ball on string, motion of 62–3
- banked corners 63–6
- base load electricity 387, 389–90
- batteries 117
- beamlines 478–9
- Bell, Alexander Graham 151–2
- bending forces 236–7
- bending losses in optical fibres 533
- Bohr, Niels
 - energy levels of hydrogen 454–5
 - spectra model for electron orbits 449–50
- booster rings 475–6
- Bragg, William Lawrence 419
 - equation of 440
 - law of 492
 - spectrometer 492
 - X-ray diffraction 491–3
- brittle materials 250–1
- candle flames in space 512
- cantilever beams 236–7, 271–2
- capacitors 288–9
 - in power smoothing 302–4
 - in R–C circuits 295
 - time-constant (τ) 294–5, 303
 - to control oscillation 317
 - types and functions 293–4
- car safety 36–7
- cathode ray oscilloscope (CRO) 134
- cathode ray tubes 465–8
- centre of mass 6
 - and movement 28
 - and stability 266–7
- centripetal acceleration 58–9
- centripetal force 59–60
- CERN 471
- circular motion
 - forces causing 59–60
 - horizontal 62–6
 - uniform 57–8
 - vertical plane 68–70, 72–3
- coefficient of friction 21
- coherent light 518
- coherent light production
 - by spontaneous emission 519
 - by stimulated absorption 519
 - by stimulated emission 519–20
- collisions, elastic and inelastic 48–9
- communications carriers, old types 539
- commutators 352
- composite materials 242–3
- compression and compressive forces 235
- Compton scattering 499–500
- Compton shift 500
- conductors, energy bands in 515
- conservation of energy 46–7
- conservation of momentum 39–42
- constructive interference of waves 412–13
- copper cables, transmission
 - frequencies and information capacity 149–50
- corners
 - banked 63–5
 - leaning into 66
- crumple zones in cars 37
- cyclotrons 470
- damping (loudspeakers) 592–3
- DC generators 375
- DC power transmission, high voltage 389–90
- de Broglie, Louis 455–6
 - wavelength 437
- decibel scale 566–8
- destructive interference (waves) 412–13
- detectors, light
 - junction 158–61
 - optical 156–8
- diffraction gratings 417
- diffraction of light 410–11
 - limits on optical instruments 416
 - observation of 414–16
- diffraction of matter waves 438–9
- diffraction of sound
 - and frequency 562–3
 - and wavelength 561–2
- diffuse scattering 501
- diodes 129–34, 287–9
 - laser (LDs) 166–8
- direct current (DC) 283–4
- doping (semiconductors) 131–2
- double-slit experiment 411–12
- ductile materials 250–1
- Earth
 - determination of mass 81
 - magnetic field of 331, 344, 347, 368
- eddy currents 369
- Einstein, Albert
 - electrodynamics of moving bodies 200–1
 - Galilean principle of relativity 197–9
 - Gedanken train 202–4, 207–10
 - Newton’s theories 10
 - space twins paradox 211–13
 - special theory of relativity 199–200, 224–5, 225–7
 - time-dilation equations 210
 - total energy equation $E = mc^2$ 222–4
- elastic collisions 48–9
- elastic modulus 249
- elastic potential energy 52–5
- electromagnets 341–3
- electric currents
 - effect of magnetic fields on 334–5
 - measurement of 125–6
- electric guitars 366
- electric motors 351–4
- electric power generators, in power stations 373–6
- electric power transmission 386–7, 389–90
- electrical safety 387
- electricity generation
 - clean 388–9
 - three-phase 379
- electrodynamics of moving bodies 200–1
- electroluminescence 514–16
- electromagnetic induction 359–60, 369
- electromagnetic spectrum 418–19
- electromagnetic waves 390–1
- electromagnetism 333–4
 - right-hand force (palm) rule 334, 347
 - right-hand grip rule 334
- electromotive force (EMF)
 - direction of 368–70
 - generated by rotating coil in magnetic field 374–5
 - magnitude 363–5

- electron guns 465–6
 electron microscopy 442
 electron scattering in crystals 439–41
 electron standing waves 455–6
 electronic circuits 117–21, 282–3
 electronic components 285
 capacitors 288–9
 diodes 287–8
 resistors 286–7
 transistors 289–90
 electronics, practical 285
 electrons
 acceleration in electric field 466–7
 charge on 443–4
 electrovolts (eV) 423–4
 emission spectra 446–7
 energy
 conservation of 46–7
 elastic potential 52–5
 forms of 44
 gravitational potential 44, 46
 kinetic 44
 and power 45–6
 of satellites in circular orbits 99
 and work 45
 energy bands in solids 515
 energy-level diagrams 450–1
 equilibrium
 rotational 269
 static 269–71
 translational 260–2

 Faraday, Michael 360
 law of induction 364–5
 fatigue in metals 258
 ferromagnetism 343
 fibre-optic sensors
 extrinsic 542
 imaging bundles in 544
 intrinsic 542–4
 in microscopy 545
 systems 541–2
 see also optical fibres
 field lines, magnetic 341
 filters, high- and low-pass 296
 fluorescent lamps 513
 force–time graphs 34–5
 forces
 bending 236–7
 causing circular motion 59–60
 centripetal 59–60
 on charged particles in magnetic fields 467–8
 on charges moving in magnetic fields 346–8
 compressive 235
 on current-carrying wires in magnetic fields 337–9
 gravitational 82
 normal 18
 shear 236
 tensile 234–5
 four-dimensional spacetime 217–18
 frames of reference 10, 188
 non-inertial 198
 Fresnel reflection 533
 friction, coefficient of 21
 full-wave rectification 300–2

 Galileo 187
 principle of relativity 188–90
 galvanometers 353, 359
 Gedanken train 202–4, 207–10
 gravitation, Newton's law of 78–81
 gravitational fields 84–5
 energy changes of moving objects in 100–3
 and falling objects 86–8
 strength of 85–6
 gravitational forces 82
 gravitational potential energy 46–7
 gravity meters (gravimeters) 87

 half-wave rectification 299–300
 harmonics 581
 harmonies 585
 hearing and sound frequency 571–2
 heat sinks 308–9
 Heisenberg, Werner 456–7
 Hertz, Heinrich 391
 high-pass filters 296
 Hooke, Robert, law of 51–2
 Hubble, Edwin 94
 Huygens, Christian 407–8
 hydrogen, energy levels of 454–5

 impulse 34–6
 incandescent light bulbs 510
 inclined planes 18–20
 incoherent light sources 508–16
 induced currents 360–1
 induced electromotive force (EMF) 363–5
 induced magnetism 328, 341–3
 induction
 electromagnetic 359–60
 Faraday's law of 364–5
 induction motors, AC 383–4
 inelastic collisions 48–9
 infrasound 573
 insertion devices 480–2
 insulators, energy bands in 515
 intensity of sound waves 565–6
 interference effects 412–13
 interference of light waves 411–12
 interference patterns
 factors affecting 413–14
 fringe spacing in 413–14
 and wavelength 414
 Internet and optical fibres 153
 invariant spacetime interval 218–19
 iron, magnetisation of 328, 341–3

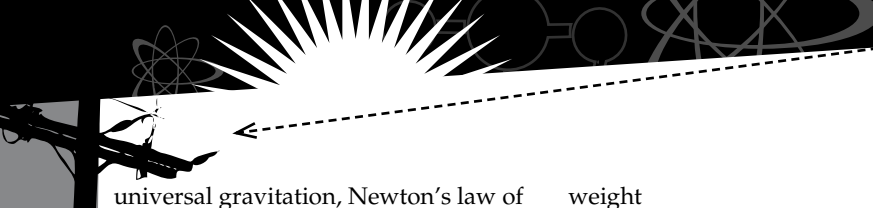
 Joule, James 45
 junction detectors 158–61

 Kepler, Johannes, laws of 95
 kinetic energy 44
 for electron release 428–9
 of photoelectrons 423

 laser diode pointers 171
 laser diodes (LDs) 166–8
 use in optical fibres 534–5
 lasers
 operation of 520–1
 safety in use 171–2
 types 521
 length contraction equation 216
 length and relative motion 215–16
 Lenz, Heinrich, law of 368
 lift-off by cars 70
 light
 absorption in optical fibres 532–3
 capture in optical fibres 523–4
 communicating with 523
 diffraction of 410–11
 dual nature of 431–2
 as electromagnetic waves 390–1
 entrapment in optical fibres 529–31
 Huygens' wave model 407–8
 interference of 411–12
 loss in optical fibres 531–4
 models of 405, 408–9
 Newton's corpuscular model 406–7
 photons of 425–6
 in spacetime 221
 speed of 190–1, 221–2
 transmission through optical fibres 525
 wave model of 410–19
 wavelength of visible 415
 light clock 208–10
 light sources
 coherent 508–9, 518–21
 continuous wide-spectrum 509–11
 incoherent 508–9, 508–16, 518
 narrow-spectrum discrete 512–13
 light-dependent resistors (LDRs) 156–8
 light-emitting diodes (LEDs) 163–5
 colours of 516
 as light source for optical fibres 534–5
 linear accelerators (linacs) 469
 in synchrotrons 474–5
 Lorentz, H.A., contraction factors 216, 217

- loudness 565
loudspeakers 591–3
 damping in 592–3
 moving coil 592
loudspeaker enclosures, baffles and ports in 593
low-pass filters 296
luge
 acceleration along 19
 coefficient of friction in 21
- maglev trains 369
magnetic attraction 328
magnetic domains 342
magnetic field
 strength of Earth's 331, 344
 unit of (tesla) 337
magnetic field lines 341
magnetic fields 327, 328–30
 around currents, magnets and atoms 341–4
 and electric currents 334–5
 forces on charged particles in 467–8
 forces on charges moving in 346–8
 forces on current-carrying wires in 337–9
 typical values 338
magnetic flux 359–60
magnetic poles 327–8
magnetism
 induced 328, 341–3
 and relativity 227–8
 simple 327–8
magnets
 natural 329
 permanent 343
 superconducting 370
mass
 centre of 6
 relativistic 222
 vs. weight 105
mass spectrometers 348–9
matter waves 437
 diffraction of 438–9
matter-scattering experiments 439–41
Maxwell, James Clerk
 conundrum of 192–3
 equations of 191, 390
 wave theory of 417
mercury-vapour lamps 451–2
metal-halide lamps 452, 513
metal-vapour lamps 512–13
 spectra of 451–2
Michelson–Morley experiment 193–5
microphones 365, 588
 electret-condenser 590–1
 electrodynamic 590
 piezoelectric 589
 velocity 590
- Millikan, Robert 443–4
modal patterns in optical fibres 535–6
modulus of elasticity 249–50
momentum 33–6
 change in 34
 conservation of 39–42
 of photons 430–1
momentum equation, relativistic 221–2
motion
 of ball on string 62–3
 describing 4–6
 equations of 8–9
 graphing 6–8
 horizontal circular 62–6
 Newton's laws 13–15
 of projectiles 23–7
 relative 10–11
 of roller-coasters 72–3
 uniform circular 57–8
 vertical 9
motors, electric 351–4
multi-dimensional worlds 201
multimeters 124, 285
muons 225–6
music
 harmonies and scales 585
 related and unrelated frequencies in 585
- Newton, Isaac 2
 assumptions of 199–200
 model of light 406–7
 law of universal gravitation 78–81
 laws of motion 13–15
 and satellite motion 90
non-inertial frames of reference 198
non-ohmic devices 129–34
normal forces 18
numerical aperture in optical fibres 530–1
- Oersted, Hans Christian 333
Ohm, Georg, law of 118
ohmic devices 129
optical detectors 156–8
optical fibre applications
 sensors 541–5
 telecommunications 538–41
optical fibres
 acceptance angle in 529–30
 advantages of 525–6
 bandwidth of 534
 bending losses in 533
 choice of types 540
 construction of 524–5
 Fresnel reflection in 533
 graded index multi-mode 527, 529
 and Internet 153
 light absorption in 532–3
- light attenuation in 531–4
light entrapment in 529–31
light sources for 534–5
material dispersion in 534
modal patterns in 535–6
numerical aperture in 530–1
optical transmitters in 534–5
Rayleigh scattering in 532
single-mode 527–8
step-index 526–7
transmission process in 525
 see also fibre optic sensors
optical instruments and diffraction 416
optical transducers 155
orbital properties of satellites,
 calculation of 93–5
- particle accelerators 348–9
 cathode ray tubes 465–8
 cyclotrons 470
 linear (linac) 469
 Van de Graaff 465
photoconductive detectors 156–8
photodiodes 158–61
photoelectric effect
 Einstein's explanation 428–30
 equation 499
 experiment 421–2
 of X-rays on metal 498–9
photoelectron ejection, threshold
 frequency for 422
photoelectron energy and stopping
 voltage 422–3
photoelectrons, kinetic energy of 423
photon energy equation 447
photon model of light 425–6
photonics 148–9
 in telecommunications 173–7
photons, momentum of 430–1
photophone, Bell's 151–2, 523
photoresistors 56–8
phototransistors 162–3
pitch of sound 571
Planck, Max 425
 constant of 426
planes, inclined 18–20
planets, data for 96
Pluto (dwarf planet) 96
Poisson bright spot 416
pole vaulting, energy changes in 54–5
position-time graphs 7
powder diffraction 441
 analysis 494–5
power
 in AC circuits 378–9
 diodes 288
 and energy 45–6
power station generators 376
power stations, electric 376, 387–8

- power supplies
 construction of 312–16
 fault-finding in 318
 hot spots in 313
 rectifiers for 299–302
 regulation of 316–18
 smoothing of 302–4, 315–16
 power transmission 386–7, 389–90
 projectile motion 23–7
 proper time and length 216–17
- radio waves 418
 radioisotopes from cyclotrons 470
 Rayleigh scattering 532
 rectification
 full-wave 300–2
 half-wave 299–300
 rectifier circuits, basic 314–15
 reflection of particles 406–7
 refraction
 of light 407–8
 of particles 406–7
 Snell's law of 405, 408
 regulation of voltage 304–9
 reinforced concrete 242–3
 relative motion 10–11
 and length 215–16
 relativistic mass 222
 relativistic momentum equation 221–2
 relativity
 Galileo's principle 188–90
 and magnetism 227–8
see also special theory of relativity
 Relenza (drug) 496
 resistance measurement 126
 resistor-capacitor (R-C) circuits 295–6
 resistors 286–7
 colour code for 287
 light-dependent 156–8
 resonance of sound 577–9
 right-hand force (palm) rule 334, 347
 right-hand grip rule 334
 RMS current and voltage 284
 rocket propulsion 41–2
 roller-coasters 72–3
 rotational equilibrium 269
- safety factors 241
 satellites
 acceleration of 93
 artificial 91–2
 calculation of orbital properties 93–5
 in circular orbits 92–3
 energy in circular orbits 99
 gravitational force on 93
 Hubble Space Telescope 96
 meteorological 96, 97
 natural 91
 in orbit 90
- SuitSat 102
 scalar quantities 3–4
 scales, musical 585
 scattering of X-rays 498–502
 semi-conductors
 doped 131
 energy bands in 515
 n-type and p-type 132–3
 shear forces 236, 237
 simultaneity and space-time 204–5
 sinusoidal current 155
 skin effect in copper cables 149–50
 skydiving and skysurfing 16
 small-angle scattering 501–2
 Snell's law of refraction 405, 408
 sodium-vapour lamps 451–2
 solar cells 160–1
 solenoids 341–2
 sonar, and whale deaths 569
 sound
 decibel scale 566–8
 diffraction of 561–3
 frequency and hearing 571–2
 frequency and pitch 571
 harmonics in 581
 as longitudinal waves 551–2
 loudness and intensity 565–8
 made visible 555
 resonance and frequency 577–9
 speed of 557–8
 transmission 550
 as transverse waves 553
 wave equation for 558–9
 wave representation of 553–4
 sound systems 588
 loudspeakers 591–3
 microphones 588–91
 spacetime 187–8
 four-dimensional 217–18
 and simultaneity 204–5
 special theory of relativity
 proofs of 224–5
 simplicity and beauty of 225–7
see also relativity
 spectroscopy with synchrotron light
 502–3
 speed 4
 average 57
 stability, and centre of mass 266–7
 standing waves 457–8
 of electrons 455–6
 sound 579–80
 static equilibrium 269–71
 stopping voltage and photoelectron
 energy 422–3
 storage rings 476–8
 strain energy 254–6
 strain in loaded materials 245–6
 strength of materials 241–2
- stress, calculation of 239–41
 stress-strain graphs 248–9
 structures in history 274–6
 struts and ties 273
 superposition of sound waves 575–6
 beats due to 576–7
 in noise reduction 577
 in synthesisers 577
 synchrotron radiation (light) 465, 473, 477, 486–7
 synchrotrons
 Australian 348–9, 464, 474–82, 483
 beamlines in 478–9
 booster ring in 475–6
 evolution of 473
 generations of 479–80
 insertion devices in 480–2
 linac in 474–5
 storage ring in 476–8
- telecommunication systems
 high fidelity 173–7
 laser-based 152–3
 networks 539–40
 temperature-dependent resistors
 (thermistors) 157
 tension and tensile forces 234–5
 Tesla, Nikola 337
 thermal radiation 511
 thermistors 157
 and voltage dividers 123–4
 thermographs 511
 Thomson's e/m experiment 443
 Thomson scattering 498
 three-phase electricity 379
 threshold frequency for photoelectron
 ejection 422
 ties and struts 273
 time
 dilation 210
 nature of 207–8
 time constant (τ) 294–5, 303
 time-varying circuits, capacitors in
 293–6
 torque 264–5
 from electric motors 352
 total energy equation $E = mc^2$ 222–4
 toughness of materials 256–7
 traffic monitors, solar-powered 161
 transformers 382–3
 equations for 383
 rated voltage of 315
 transistor amplifier circuits 140–1
 transistors 289–90
 translational equilibrium 260–2
 tungsten-halogen globes 510
- ultrasound 573
 undulators 480–1



universal gravitation, Newton's law of
78–81

Van de Graaff particle accelerator 465

vector quantities 3–6

velocity 4

velocity–time graphs 7

vertical motion 9

voltage

measurement 125

regulation of 304–6

voltage amplifiers 137

voltage dividers 121–2

voltage dividing, and thermistors
123–4

voltage gain 138

voltage regulators 306–9

voltmeters, analogue 353

watt hour meters 388

wave equation 404–5

wave model of light 410

inadequacy of 424–5

wavelength 404

de Broglie 437

visible light 415

waves 403–4

amplitude of 565

and energy transfer 403

frequency of 403–4

interference effects of 412–13

path difference of 412

period of 403–4

phase 508

radio 418

standing 579–80

weight

apparent 106

in gravitational field 85

vs. mass 105

weightlessness 107–8

whale deaths and sonar 569

white-light-emitting diodes (WLEDs)
165–6

Wien, Wilhelm, law of 511

wigglers 480–1

wind instruments

with closed air columns 583–4

with open-ended air columns 582–3

work and energy 45

work function (w) for release of
electron from metal 428

work hardening of metals 258

X-ray diffraction 491–3

uses of 493–4

X-ray lithography 490

X-rays 418–19

and matter interaction 498–502

properties of 488–9

scattering in crystals 439

Young, Thomas, double-slit

experiment 411–12

Young's modulus (E) 249–50

zener diodes 288