HL

PEARSON BACCALAUREATE

HIGHER LEVEL

Physics

DEVELOPED SPECIFICALLY FOR THE **IB DIPLOMA**

CHRIS HAMPER

With integrated online learning resources!
Visit www.pearsonbacconline.com

ALWAYS LEARNING

PEARSON

PEARSON
BACCALAUREATE

HIGHER LEVEL (plus STANDARD LEVEL OPTIONS)

Physics

DEVELOPED SPECIFICALLY FOR THE

IB DIPLOMA

CHRIS HAMPER

PEARSON

Pearson Education Limited is a company incorporated in England and Wales, having its registered office at Edinburgh Gate, Harlow, Essex, CM20 2JE. Registered company number: 872828

www.pearsonbaccalaureate.com

First published 2009

13 12 11 10
10 9 8 7 6 5 4

ISBN 978 0 435994 42 6

Edited by Maggie Rumble
Designed by Tony Richardson
Typeset by Tech-Set Ltd
Original illustrations © Pearson Education Limited 2009
Illustrated by Tech-Set Ltd
Picture research by Joanne Forrest Smith
Cover design by Tony Richardson
Cover photo © Science Photo Library/Alfred Pasieka
Printed by Rotolito in Italy

Acknowledgements and disclaimers
The author and publisher would like to thank Keith Ord, the co-author of Pearson Baccalaureate Physics Standard Level, on which parts of this book are based; Joonas Govenius for the abstract on page 576 and the following individuals and organisations for permission to reproduce photographs:
p.4 © Corbis / Martial Trezzini; p.18 ImageShow-POD; p.28 © SPL/ Edward Kinsman; p.30 © SPL/ NASA; p.33 © SPL/ Philippe Plailly; p.59 © Corbis / Peter Dench; p.67 © SPL/ Charles D Winters; p.68 © SPL/ Mehau Kulyk; p.69 © SPL/ Edward Kinsman; p.72 © SPL/ Dr Arthur Tucker; p.74 © SPL/ John Beatty; p.79 © SPL/ Charles D Winters; p.100 © Digital Vision; pp.110,112 © SPL/ Andrew Lambert Photography; p.113 © iStockPhoto.com / Paul Piebinga; p.115 © SPL/ FS Westmoreland; p.116 © SPL/ Martin Bond; p.119 © SPL/ David Weintraub; p.122 © Getty Images / PhotoDisc; p.123 © SPL/ Alex Bartel; p.125 © SPL/ Erich Schrempp; p.126 © SPL/ Andrew Lambert Photography; p.127 © SPL/ Edward Kinsman; p.128 © SPL/ David Parker; p.129 © SPL/ Pasquale Sorrentino; p.131 © Alamy; p.134 © Alamy; p.144 © SPL/ Cordelia Molloy; p.147 © SPL/ Peter Aprahamian / Sharples Stress Engineers Ltd; p.148 © Getty Images / PhotoDisc; p.157 © Pearson Education Ltd / Trevor Clifford; p.160 © SPL/ Sheila Terry; p.161 © SPL/ Andrew Lambert Photography; p.163 © SPL/ David R Frazier; p.170 © Getty Images / PhotoDisc; p.181 © Pasco; p.183 © Getty Images / PhotoDisc; p.192 © Corbis; p.194 SPL/NASA; p.204 tl. © SPL/ Andrew Lambert Photography; p.204 br. © Alamy Images / Tetra; p.233 © Digital Vision; p.242 © SPL/ Mehau Kulyk; p.245 © SPL; p.246 © SPL / N Feather; p.247 © SPL; p.248 SPL/Lawrence Berkeley Laboratory; p.250 © SPL/ Maria Platt-Evans; p.251 © SPL/ Philippe Psaila; p.252 © SPL/ Prof. K. Seddon & Dr T Evans, Queens Univ. Belfast; p.257 © SPL/ Seymour; p.268 © SPL/ Maximillion Stock Ltd; p.272 © Brand X Pictures; p.275 t. © SPL/ Simon Fraser; p.275 b. © Digital Vision; p.279 © SPL/ Ria Novosti; p.284 © SPL/ EFDA-JET; p.286 © SPL/ Detlev Van Ravenswaay; p.287 © SPL/ Sheila Terry; p.288 © SPL/ Bernhard Edmaier; p.289 © SPL/ Martin Bond; p.291 © Punchstock / Fancy; p.303 © SPL/ John Chumack; p.305 © SPL/ Dr Juerg Alean; p.312 © SPL/ Dr Tony Brain; p.317 © SPL/ Christian Darkin; p.318 © SPL / Dr Jeremy Burgess; p.321© SPL / Detlev Van Ravenswaay; p.324 © Pearson Education Ltd / Gareth Boden; p.325 © SPL / Thomas Porett; p.327 © Isabell Eischeid; p.329 © Getty Images / PhotoDisc; p.332 SPL/Skyscan; p.333 © SPL / Ron Sandford; p.335 Corbis; p.336 Corbis; p.337 t. © SPL / Ronald Reyer; p.337 b. © SPL/John Sanford; p.338 © SPL / Mark Garlick; p.343 Christian Darkin / SPL and National Optical Astronomy Observatories / SPL; p.355 European Southern Observatory / SPL; p.356 Royal Observatory Edinburgh / SPL; p.357 Mark Garlick / SPL; p.359 Royal Observatory Edinburgh / AATB / SPL; p.369 © SPL/ David May Jones; p.384 t. © Alamy Images / SciencePhotos; p.384 m. © Harcourt Ltd / Trevor Clifford; p.384 b. © Alamy Images / M Shields Photos; p.386 © Alamy Images; p.395 © SPL; p.397 © Omegatron; p.403 © SPL / Library of Congress; p.407 © SPL / Colin Cuthbert; p.411 © SPL / Sheila Terry; p.415 © SPL/ Mehau Kulyk; p.416 t. © SPL/ David Parker; p.416 b. © Alamy Images; p.417 © SPL/ Sheila Terry; p.418 © Getty Images / PhotoDisc; p.419 © SPL/ R Mesna, Fundamental Photo; p.420 © SPL/ Coneyl Jay; p.428 © SPL / John Heseltine; p.430 © SPL / Dr Jeremy Burgess; p.435 © SPL / Javier Trueba / MSF; p.436 © SPL; p.437 © Corbis; p.438 © Pearson Education Ltd / Tudor Photography; p.444 © Corbis; p.469 © Corbis; p.475 t. © SPL / Royal Astronomical Society; p.477 b. © SPL / A Bolton / H-S CfA / SLACS / NASA / ESA / STScI; p.486 © SPL/ William Ervin; p.487 © SPL / Camilla Elam; p.492 t. © SPL/ Martyn F. Chillmaid; p.492 b. Chris Hamper; p.495 t. ©SPL / Gustoimages; p.495 b. ©SPL / Antoine Rosset; p.498 © SPL / Dr Najeeb Layyous; p.499 © iStockPhoto; p.500 © SPL / MIT AI Lab / Surgical Planning Lab / Brigham & Womens Hospital; p.501 © SPL / AJ Photo / HOP American; p.507 t. ©SPL / CNRI; p.507 b. SPL / Wellcome Department of Cognitive Neurology; p.513 © SPL / Andrew Lambert Photography; p.520 © SPL / Goronwy Tudor Jones, University of Birmingham; p.528 t. © SPL / CERN / P Loiez; p.528 b. © SPL / PIP Kalmus, UAI Experiment; p.529 © SPL / Stanford Linear Accellerator Centre; p.530 © SPL / PIP Kalmus, UAI Experiment; p.550 t. © SPL/ Ria Novosti; p.550 br. © Getty Images / PhotoDisc; p.550 bl. © SPL/ Dr Matk J Winter; p.551 t. © SPL/ Sheila Terry; p.551 bl. © SPL; p.551 br. © iStockPhoto.com / Lise Gagne; p.552 © Jupiter Images / Photos.com; p.553 © SPL; p.554 t. © Mike Baldwin / Cornered; p.554 b. © Digital Vision; p.555 © Digital Vision; p.556 t. © SPL/ B McNamara (University of Waterloo), NASA, ESA, STSci; p.556 m. Erik Fendik; p.556 b. © SPL/ KH Kjeldsen; p.557 t. © Digital Vision; p.557 b. © Harcourt Ltd / Tudor Photography; p.573 Chris Hamper / Giovanni Braghieri; pp.577, 580 © SPL / David Parker.

Websites
There are links to relevant websites in this book. In order to ensure that the links are up-to-date, that the links work, and that the sites are not inadvertently linked to sites that could be considered offensive, we have made the links available on our website at www.heinemann.co.uk/hotlinks. When you access the site, the express code is 4426P.

Dedication

There are several people I would like to dedicate this book to, firstly my father who persuaded me to take up teaching, my family who supported me through the writing process and all the students I have ever taught whose questions came back to me throughout the lonely hours of writing. There were also five students, Bante, Liuliang, Xiaolong, DANAMONA and Giovanni who deserve special thanks for their help in proofing the problem solutions.

Chris Hamper

Contents

Introduction

Welcome to your new course! This book is designed to act as a comprehensive course book, covering the core and AHL material and all the options you might take while studying for the IB Diploma in Physics at Higher Level. It will also help you to prepare for your examinations in a thorough and methodical way.

Content

Chapters 1 to 9 provide coverage of the core and AHL material. Chapters 10 to 15 cover the options E to J. HL students are required to study two of these options. Within each chapter, there are numbered exercises for you to practise and apply the knowledge that you have gained. The exercises will also help you to assess your progress. Sometimes, there are worked examples that show you how to tackle a particularly tricky or awkward question.

Worked example

A car travelling at 30 m s^{-1} emits a sound of frequency 500 Hz. Calculate the frequency of the sound measured by an observer in front of the car.

Solution

This is an example where the source is moving relative to the medium and the observer is stationary relative to the medium. To calculate the observed frequency, we use the equation

$$f_1 = \frac{cf_0}{(c - v)}$$

where

$$f_0 = 500\ \text{Hz}$$

$$c = 330\ \text{m s}^{-1}$$

$$v = 30\ \text{m s}^{-1}$$

So

$$f_1 = \frac{330 \times 500}{330 - 30} = 550\ \text{Hz}$$

At the end of each chapter, there are practice questions taken from past exam papers. Towards the end of the book, starting on page 582, you will find answers to all the exercises and practice questions.

After the options chapters, you will find a Theory of Knowledge chapter, which should stimulate wider research and the consideration of moral and ethical issues in the field of physics. Following this, there are two short chapters offering advice on internal assessment and on writing extended essays.

Finally, there is a short appendix which covers material on the eye and sight. This is not required study for HL students, but is part of Option A for SL students.

Information boxes

You will see a number of coloured boxes interspersed through each chapter. Each of these boxes provides different information and stimuli as follows.

Assessment statements

7.3.6 Draw and annotate a graph showing the variation with nucleon number of the binding energy per nucleon.

7.3.7 Solve problems involving mass defect and binding energy.

13.2.2 Describe how the masses of nuclei may be determined using a Bainbridge mass spectrometer.

13.2.1 Explain how the radii of nuclei may be estimated from charged particle scattering experiments.

You will find a box like this at the start of each section in each chapter. They are the numbered objectives for the section you are about to read and they set out the content and aspects of learning covered in that section.

Infinity
We can't really take a mass from infinity and bring it to the point in question, but we can calculate how much work would be required if we did. Is it OK to calculate something we can never do?

In addition to the Theory of Knowledge chapter, there are TOK boxes throughout the book. These boxes are there to stimulate thought and consideration of any TOK issues as they arise and in context. Often they will just contain a question to stimulate your own thoughts and discussion.

Antimatter
It is also possible for a proton to change into a neutron and a positive electron plus a neutrino. The positron is in some way the opposite of an electron – it is called an *antiparticle*. The antineutrino $\bar{\nu}$ is also the antiparticle of the neutrino. Every particle has an antiparticle. Atoms made of negative positrons and positive electrons are called *antimatter*.

These boxes contain interesting information which will add to your wider knowledge but which does not fit within the main body of the text.

The electronvolt
Remember 1 eV is the KE gained by an electron accelerated through a p.d. of 1 V.

$$1\ \text{eV} = 1.6 \times 10^{-19}\ \text{J}$$

These facts are drawn out of the main text and are highlighted. This makes them useful for quick reference and they also enable you to identify the core learning points within a section.

Different cultures developed different systems for defining the different patterns in the stars. The Greeks had the 12 signs of the zodiac, the Chinese 28 xiu and the Indians 27 Nakshatra.

These boxes indicate examples of internationalism within the area of study. The information in these boxes gives you the chance to think about how physics fits into the global landscape. They also cover environmental and political issues raised by your subject.

- **Examiner's hint:** Since field strength g is a vector, the resultant field strength equals the vector sum.

These boxes can be found alongside questions, exercises and worked examples and they provide insight into how to answer a question in order to achieve the highest marks in an examination. They also identify common pitfalls when answering such questions and suggest approaches that examiners like to see.

To view a simulation of random decay, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 7.4.

These boxes direct you to the Heinemann website, which in turn will take you to the relevant website(s). On the web pages you will find background information to support the topic, or video simulations and animations.

Using this book for SL Physics

This book has been written for students studying HL physics but can also be used by SL students. If you are studying SL physics you do not have to study all of the HL core material.

SL options

SL options A, B and C are the same as the HL core:
A Sight and Wave Phenomena
This is the same as the AHL material in Chapter 4 plus the material on the eye in the Appendix (pp. 577–581).

B Quantum and Nuclear Physics
This is the same as the AHL material in Chapter 7.

C Digital Technology
This is the same as the AHL material in Chapter 9 plus the section on electronics and the mobile phone in Chapter 11.

So if you are an SL student and you study up to and including Chapter 9 you will have done the SL core plus three options (you only need to do two).

The remaining SL options are parts of the HL options.

D Relativity and Particles
This is the first part of Chapter 13 and the first part of Chapter 15.

E Astrophysics
This is the first part of Chapter 10.

F Communications
This is the same as Chapter 11 without the part on electronics and the mobile phone.

G Electromagnetic Waves
This is the first part of Chapter 12.

Worked Solutions

Full worked solutions to all exercises and practice questions can be found online at www.pearsonbacc.com/solutions.

1 Physics and physical measurement

1.1 Fundamental quantities

Assessment statements

1.1.1 State and compare quantities to the nearest order of magnitude.
1.1.2 State the ranges of magnitude of distances, masses and times that occur in the universe, from the smallest to the greatest.
1.1.3 State ratios of quantities as differences of orders of magnitude.
1.1.4 Estimate approximate values of everyday quantities to one or two significant figures and/or to the nearest order of magnitude.
1.2.1 State the fundamental units in the SI system.
1.2.2 Distinguish between fundamental and derived units and give examples of derived units.
1.2.3 Convert between different units of quantities.
1.2.4 State units in the accepted SI format.
1.2.5 State values in scientific notation and in multiples of units with appropriate prefixes.

Position

Physics is about modelling our universe. To do this, we need to define the things inside it. Each thing is different for many reasons, but one of the most important differences is their different positions. To define position, we use the quantity *distance*; this is how far the object is away from us. To quantify (put a number to) this difference we compare the distance with some standard measure (the metre rule). All distances can then be quoted as multiples of this *fundamental unit*, for example:

The distance from Earth to the Sun $= 1.5 \times 10^{11}$ m
The size of a grain of sand $= 2 \times 10^{-4}$ m
The distance to the nearest star $= 4 \times 10^{16}$ m
The radius of the Earth $= 6.378 \times 10^{6}$ m

Standard form

The quantities above are expressed in standard form. This means that there is only one number to the left of the decimal place. For example:
1600 m in standard form is 1.6×10^{3} m

Exercise

1 Convert the following into metres (m) and write in standard form:
(a) Distance from London to New York = 5585 km
(b) Height of Einstein was 175 cm
(c) Thickness of a human hair = 25.4 μm
(d) Distance to edge of the universe = 100 000 million million million km

The metre
The metre was originally defined in terms of several pieces of metal positioned around Paris. This wasn't very accurate so now one metre is defined as the distance travelled by light in a vacuum in $\frac{1}{299\,792\,458}$ of a second.

It is also acceptable to use a prefix to denote powers of 10

Prefix	Value
T (tera)	10^{12}
G (giga)	10^{9}
M (mega)	10^{6}
k (kilo)	10^{3}
c (centi)	10^{-2}
m (milli)	10^{-3}
μ (micro)	10^{-6}
n (nano)	10^{-9}
p (pico)	10^{-12}
f (femto)	10^{-15}

The second
The second was originally defined as a fraction of a day but today's definition is 'the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium-133 atom'.

If nothing ever happened, would there be time?

The kilogram
The kilogram is the only fundamental quantity that is still based on an object kept in Paris. Moves are underway to change the definition to something that is more constant and better defined but does it really matter? Would anything change if the size of the 'Paris mass' changed?

Time

When something happens we call it an *event*. To distinguish between different events we use time. The time between two events is measured by comparing to some fixed value, the second. Time is also a fundamental quantity.

Some examples of times:
Time between beats of a human heart = 1 s
Time for the Earth to go around the Sun = 1 year
Time for the Moon to go around the Earth = 1 month

Exercise

2 Convert the following times into seconds (s) and write in standard form:
(a) 85 years, how long Newton lived
(b) 2.5 ms, the time taken for a mosquito's wing to go up and down
(c) 4 days, the time it took to travel to the Moon
(d) 2 hours 52 min 59 s, the time for Concord to fly from London to New York

Mass

If we pick different things up we find another difference. Some things are easy to lift up and others are difficult. This seems to be related to how much matter the objects consist of. To quantify this we define mass measured by comparing different objects to a piece of metal in Paris, the standard kilogram.

Some examples of mass:
Approximate mass of a man = 75 kg
Mass of the Earth = 5.97×10^{24} kg
Mass of the Sun = 1.98×10^{30} kg

Exercise

3 Convert the following masses to kilograms (kg) and write in standard form:
(a) The mass of an apple = 200 g
(b) The mass of a grain of sand = 0.00001 g
(c) The mass of a family car = 2 tonnes

Volume

The space taken up by an object is defined by the volume. Volume is measured in cubic metres (m^3). Volume is not a fundamental unit since it can be split into smaller units ($m \times m \times m$). We call units like this *derived units*.

Exercises

4 Calculate the volume of a room of length 5 m, width 10 m and height 3 m.

5 Using the information from page 1, calculate:
(a) the volume of a human hair of length 20 cm
(b) the volume of the Earth.

Density

By measuring the mass and volume of many different objects we find that if the objects are made of the same material, the ratio mass/volume is the same. This quantity is called the *density*. The unit of density is $\mathrm{kg\,m^{-3}}$. This is another derived unit.

Examples include:
Density of water $= 1.0 \times 10^3\,\mathrm{kg\,m^{-3}}$
Density of air $= 1.2\,\mathrm{kg\,m^{-3}}$
Density of gold $= 1.93 \times 10^4\,\mathrm{kg\,m^{-3}}$

Exercises

6 Calculate the mass of air in a room of length 5 m, width 10 m and height 3 m.

7 Calculate the mass of a gold bar of length 30 cm, width 15 cm and height 10 cm.

8 Calculate the average density of the Earth.

Displacement

So far all that we have modelled is the position of things and when events take place, but what if something moves from one place to another? To describe the movement of a body, we define the quantity *displacement*. This is the distance moved in a particular direction.

The unit of displacement is the same as length: the metre.

Example:
Refering to the map in Figure 1.1:
If you move from B to C,
your displacement will be 5 km north.
If you move from A to B,
your displacement will be 4 km west.

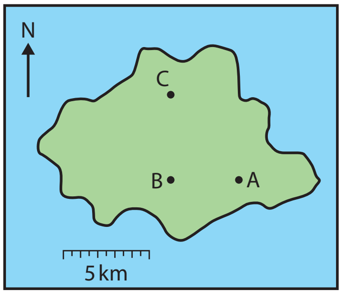


Figure 1.1 Displacements on a map.

Summary of SI units

The International System of units is the set of units that are internationally agreed to be used in science. It is still OK to use other systems in everyday life (miles, pounds, Fahrenheit) but in science we must always use SI. There are seven fundamental quantities.

Base quantity	Name	Symbol
length	metre	m
mass	kilogram	kg
time	second	s
electric current	ampere	A
thermodynamic temperature	kelvin	K
amount of substance	mole	mol
luminous intensity	candela	cd

All other SI units are derived units; these are based on the fundamental units and will be introduced and defined where relevant. So far we have come across just two.

Derived quantity	Symbol	Base units
volume	m^3	$m \times m \times m$
density	$kg\,m^{-3}$	$\frac{kg}{m \times m \times m}$

1.2 Measurement

Assessment statements

1.2.6 Describe and give examples of random and systematic errors.
1.2.7 Distinguish between precision and accuracy.
1.2.8 Explain how the effects of random errors may be reduced.
1.2.9 Calculate quantities and results of calculations to the appropriate number of significant figures.

Even this huge device at CERN has uncertainties.

Uncertainty and error in measurement

The SI system of units is defined so that we all use the same sized units when building our models of the physical world. However, before we can understand the relationship between different quantities, we must measure how big they are. To make measurements we use a variety of instruments. To measure length, we can use a ruler and to measure time, a clock. If our findings are to be trusted, then our measurements must be accurate, and the accuracy of our measurement depends on the instrument used and how we use it. Consider the following examples.

Estimating uncertainty
When using a scale such as a ruler the uncertainty in the reading is $\frac{1}{2}$ of the smallest division. In this case the smallest division is 1 mm so the uncertainty is 0.5 mm.

Measuring length using a ruler

Example 1

A good straight ruler marked in mm is used to measure the length of a rectangular piece of paper as in Figure 1.2.

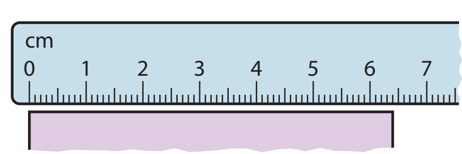


Figure 1.2 Length = 6.40 ±0.05 cm.

The ruler measures to within 0.5 mm (we call this the uncertainty in the measurement) so the length in cm is quoted to 2 dp. This measurement is precise and accurate.

Example 2

Figure 1.3 shows how a ruler with a broken end is used to measure the length of the same piece of paper. When using the ruler, you fail to notice the end is broken and think that the 0.5 cm mark is the zero mark.

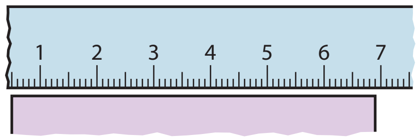


Figure 1.3 Length = 6.90 ± 0.05 cm.

This measurement is precise since the uncertainty is small but is not accurate since the value 6.90 cm is wrong.

If you measure the same thing many times and get the same value, then the measurement is precise.
If the measured value is close to the expected, then the measurement is accurate. If a football player hit the post 10 times in a row when trying to score a goal, you could say the shots are precise but not accurate.

Example 3

A cheap ruler marked only in $\frac{1}{2}$ cm is used to measure the length of the paper as in Figure 1.4.

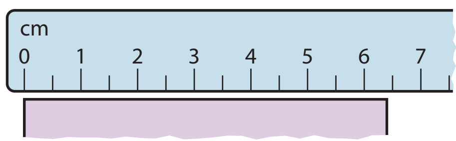


Figure 1.4 Length = 6.5 ± 0.25 cm.

These measurements are not precise but accurate, since you would get the same value every time.

Example 4

In Figure 1.5, a good ruler is used to measure the maximum height of a bouncing ball. Even though the ruler is good it is very difficult to measure the height of the bouncing ball. Even though you can use the scale to within 0.5 mm, the results are not precise (may be about ± 0.2 cm). However, if you do enough runs of the same experiment, your final answer could be accurate.

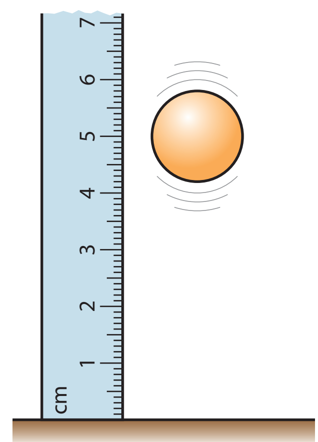


Figure 1.5 Height = 5.0 ± 0.2 cm.

Errors in measurement

There are two types of measurement error – random and systematic.

Random error

If you measure a quantity many times and get lots of slightly different readings then this called a *random error*. For example, when measuring the bounce of a ball it is very difficult to get the same value every time even if the ball is doing the same thing.

Systematic error

This is when there is something wrong with the measuring device or method. Using a ruler with a broken end can lead to a ‘zero error’ as in Example 2 above. Even with no random error in the results, you’d still get the wrong answer.

It is not possible to measure anything exactly. This is not because our instruments are not exact enough but because the quantities themselves do not exist as exact quantities.

Reducing errors

To reduce random errors you can repeat your measurements. If the uncertainty is truly random, they will lie either side of the true reading and the mean of these values will be close to the actual value. To reduce a systematic error you need to find out what is causing it and correct your measurements accordingly. A systematic error is not easy to spot by looking at the measurements, but is sometimes apparent when you look at the graph of your results or the final calculated value.

1.3 Collecting data

Assessment statements
1.2.10 State uncertainties as absolute, fractional and percentage uncertainties.
1.2.11 Determine the uncertainties in results.

Measurement in practice

Approximately one quarter of this course will be taken up with practical work, where you will be measuring quantities and doing calculations. There are certain accepted ways of handling data and the uncertainties they contain. In the following section we will go through the way to make tables, draw graphs and handle uncertainties. If you follow this method you will gain good marks in the internally assessed part of the course.

To illustrate the method we will consider a simple experiment. You are presented with 6 metal cubes and asked to find the density of the metal. The apparatus supplied is a good ruler and an electronic balance. The ruler has millimetre divisions and the balance measures down to 0.1 g.

Density is defined as mass/volume, so to find density we must measure mass and volume. As the samples are cubes we only need to measure one side. We can then find the volume by cubing this value.

Repeating measurements

Each cube has a different side length and mass. To make sure each mass is connected to the correct side length we put our data into a table; it is also simpler if we use a spreadsheet to perform calculations.

The cubes are not perfectly uniform so the side length depends upon which length we choose. To reduce the uncertainty in this measurement we measure the cube four times and find the average side length. However, if we measure the mass of the cube we get exactly the same measurement each time and there is therefore no point in repeating this measurement.

- **Examiner's hint:** It's OK to use non-SI units such as grams when collecting data. However, your final result (density) should be in SI units.

- **Examiner's hint:** The number of decimal places in the data should be consistent with the uncertainty. It would be wrong to write 2.4000 cm since the uncertainty is ± 0.05 cm

Table 1.1

Mass/g ±0.1 g	Length 1/cm ±0.05 cm	Length 2/cm ±0.05 cm	Length 3/cm ±0.05 cm	Length 4/cm ±0.05 cm
124.1	2.40	2.30	2.50	2.40
235.2	3.00	3.10	2.90	3.00
344.0	3.40	3.30	3.40	3.50
463.2	3.70	3.80	3.60	3.70
571.2	4.00	4.10	3.90	4.00
660.0	4.20	4.30	4.10	4.20

Note: The uncertainty in each length measurement is 0.05 cm. However the actual uncertainty is greater as the spread of values demonstrates.

Processing data

In this example we are first going to find the average side length and then the volume. If we look at the data we can see that the spread of data is more than the uncertainty in the ruler scale so we will also calculate the uncertainty in these values. When we know that, we can find out the uncertainty in the volume. This is found by calculating the maximum and minimum volumes using the maximum and minimum values for length. Even in this simple experiment there will be a lot of calculations, so it is advisable to use a spreadsheet (such as Excel) for doing this.

● **Examiner's hint:** It is well worth learning how to use a spreadsheet programme for analysing data.

This is calculated using the formula $((\text{max length})^3 - (\text{min length})^3)/2$

This is found by calculating (max length − (min length)/2

● **Examiner's hint:** The number of decimal places in the data must not exceed the uncertainty and the uncertainty has been rounded off to 1 significant figure.

Table 1.2

Mass/g ±0.1 g	Mean length/cm	Uncertainty in length/±cm	Volume/cm^3	Uncertainty in volume/±cm^3
124.1	2.4	0.1	14	2
235.2	3.0	0.1	27	3
344.0	3.4	0.1	39	3
463.2	3.7	0.1	51	4
571.2	4.0	0.1	64	5
660.0	4.2	0.1	74	5

1.4 Presenting processed data

Assessment statements

1.2.12 Identify uncertainties as error bars in graphs.

1.2.13 State random uncertainty as an uncertainty range (±) and represent it graphically as an 'error bar'.

1.2.14 Determine the uncertainties in the gradient and intercepts of a straight-line graph.

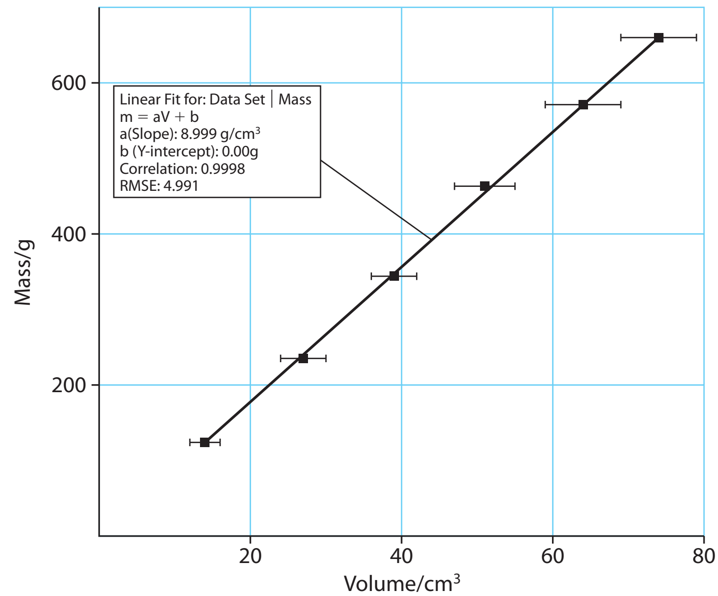
You could simply calculate the density of the metal for each cube and find the average but there are many advantages to using a graphical method. In this example, we know that the mass and volume are related by the equation $m = \rho V$ where ρ is the density.

This means that mass is proportional to volume, so plotting a graph of mass (on the y-axis) against volume (on the x-axis) will give a straight line. The gradient of the line will be the density and the y-intercept will be zero. Because there are uncertainties in the data, we don't know exactly where to plot the line; for this reason we plot error bars on each point as shown in Figure 1.6. This graph has been plotted with a computer program (e.g. Graphical Analysis by Vernier) that automatically plots the best fit line and the error bars.

The general equation of a straight line is
$y = mx + c$
Where m is the gradient and c is the y-intercept. Any equation that has the same form will also be linear.

● **Examiner's hint:** As with spreadsheets, it is also worth practising with graph plotting software.

Figure 1.6 Graph of mass vs volume with error bars
As we can see, the gradient or slope of the line is 8.999 $g\,cm^{-3}$; this is the density of the metal.

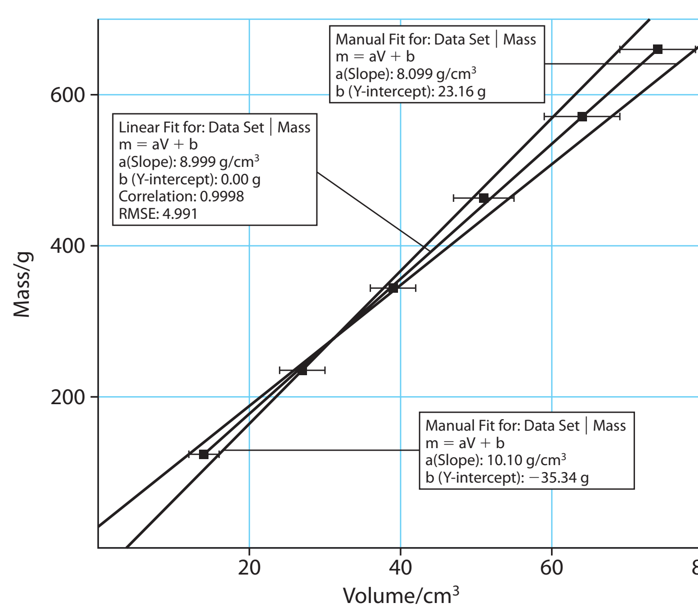


To find out more about graph plotting visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 1.1.

Uncertainties in gradients

We can see from Figure 1.6 that the line shown is not the only straight line that can be drawn through the error bars; there are in fact a whole range of them. Using the steepest and least steep lines that we can draw will give us the uncertainty in the gradient. The computer program also enables us to do that, as shown in Figure 1.7.

Figure 1.7 Graph of mass vs volume with steepest and least steep lines.



The maximum gradient is 10.1 and the minimum is 8.1 so the uncertainty is

$$\frac{(10.1 - 8.1)}{2} = 1\,g\,cm^{-3}$$

So our final result is that the density of the metal in SI units is $9000 \pm 1000\,kg\,m^{-3}$.

Note the number of significant figures is reduced so that it is consistent with the uncertainty.

If we look up the density of metals we find that the density of copper is 8920 kg m^{-3}. This is only 80 kg m^{-3} less than our value. We can therefore conclude that within the uncertainties of our experiment the cubes could be made of copper.

Percentage uncertainties

In the example above, the uncertainties were expressed as ± 1000 kg m^{-3}. This is called an *absolute uncertainty*. Uncertainties can also be expressed as a simple percentage. In this case the percentage uncertainty would be $\left(\frac{1000}{9000}\right) \times 100 = 11\%$

When you multiply values, you can find the uncertainty of the result by adding the percentage uncertainties. However when dealing with tables of data, it is simpler to use the method described previously. So if the uncertainty in length is 4%, the uncertainty in volume is $3 \times 4\% = 12\%$.

● **Examiner's hint:** Even though you use absolute uncertainties in all your practical work you might be asked how to manipulate percentage uncertainties in the exam.

Exercise

9 To measure the volume of an object, two lengths l_1 and l_2 are measured.

$l_1 = 10.25 \pm 0.05$ cm

$l_2 = 15.45 \pm 0.05$ cm

Calculate:

(a) the % uncertainty in l_1

(b) the % uncertainty in l_2

(c) the area of the object

(d) the % uncertainty in the area.

Other information from graphs

A graph is not only a way to find a value, it also gives us information about the validity of the data. Here are some examples.

An outlier

If you have made a mistake, it will show on the graph as an *outlier*. This is a point that does not fit the line, like the one in Figure 1.8.

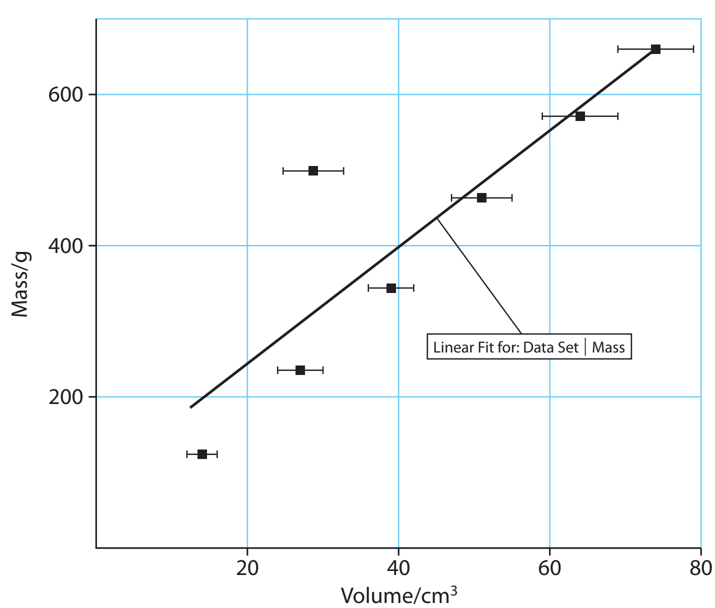


Figure 1.8 Graph with an outlier.

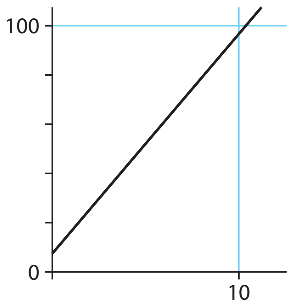


Figure 1.9 A non-zero intercept.

A non-zero intercept

If the intercept is supposed to be zero but isn't, this could be due to a systematic error. For example, if all the mass measurements had been 10 g too big, we would get an intercept of 10 g as in Figure 1.9.

A non-linear trend

Relationships in physics are not always linear. This is shown by the graph in Figure 1.10, where the straight line does not fit, but there is a clear relationship.

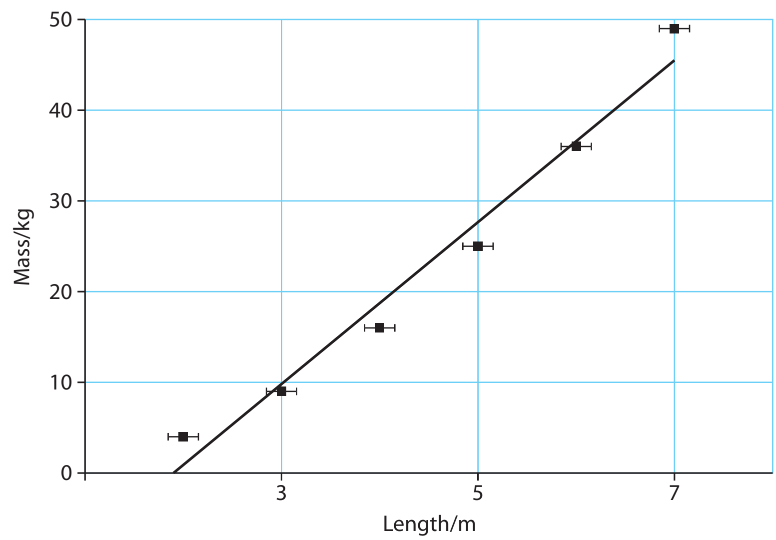


Figure 1.10 Graph with a non-linear trend – the line does not pass through the error bars.

1.5 Vectors and scalars

Scalar
A quantity with magnitude only.
Vector
A quantity with magnitude and direction.

Assessment statements

1.3.1 Distinguish between vector and scalar quantities, and give examples of each.
1.3.2 Determine the sum or difference of two vectors by a graphical method.
1.3.3 Resolve vectors into perpendicular components along chosen axes.

So far we have dealt with six different quantities:

- Length
- Time
- Mass
- Volume
- Density
- Displacement

All of these quantities have a size, but displacement also has a direction. Quantities that have size and direction are *vectors* and those with only size are *scalars*; all quantities are either vectors or scalars. It will be apparent why it is important to make this distinction when we add displacements together.

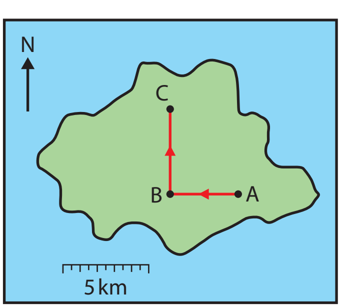


Figure 1.11 Displacements shown on a map.

Example

Consider two displacements one after another as shown in Figure 1.11.
Starting from A walk 4 km west to B, then 5 km north to C.

The total displacement from the start is not 5 + 4 but can be found by drawing a line from A to C.

We will find that there are many other vector quantities that can be added in the same way.

Addition of vectors

Vectors can be represented by drawing arrows. The length of the arrow is proportional to the magnitude of the quantity and the direction of the arrow is the direction of the quantity.

To add vectors the arrows are simply arranged so that the point of one touches the tail of the other. The resultant vector is found by drawing a line joining the free tail to the free point.

Example

Figure 1.11 is a map illustrating the different displacements. We can represent the displacements by the vectors in Figure 1.12.

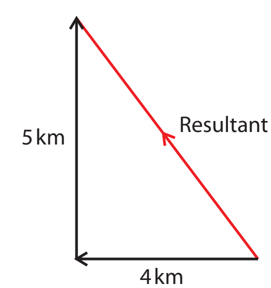


Figure 1.12 Vector addition.

Calculating the resultant:
If the two vectors are at right angles to each other then the resultant will be the hypotenuse of a right-angled triangle. This means that we can use simple trigonometry to relate the different sides.

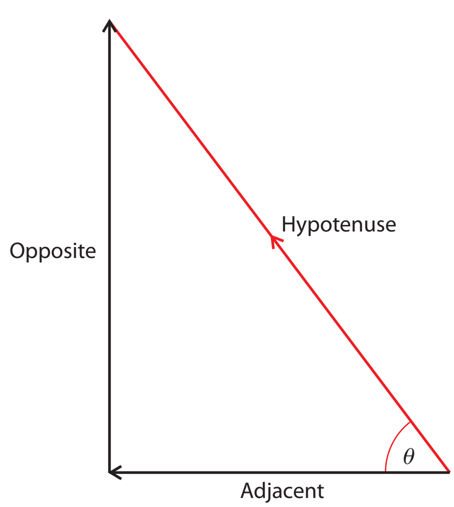


Figure 1.13

● **Examiner's hint:** Not all vectors add up to give right-angled triangles but as these are the easy ones to solve we will consider only these in this course.

You will find *cos*, *sin* and *tan* buttons on your calculator. These are used to calculate unknown sides of right-angled triangles.

$$\text{Sin } \theta = \frac{\text{Opposite}}{\text{Hypotenuse}}$$

$$\text{Cos } \theta = \frac{\text{Adjacent}}{\text{Hypotenuse}}$$

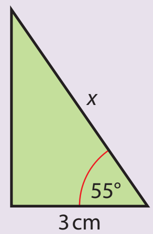
$$\text{Tan } \theta = \frac{\text{Opposite}}{\text{Adjacent}}$$

Vector symbols
To show that a quantity is a vector we can write it in a special way. In textbooks this is often in bold (**A**) but when you write you can put an arrow on the top. In physics texts the vector notation is often left out. This is because if we know that the symbol represents a displacement, then we know it is a vector and don't need the vector notation to remind us.

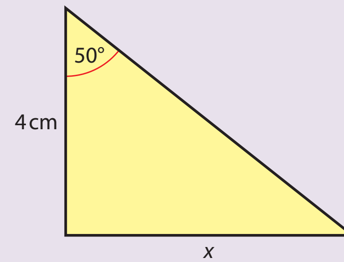
Exercise

10 Use your calculator to find x in the following triangles.

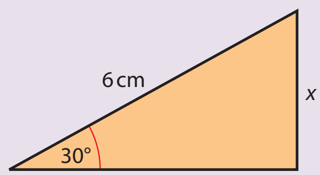
(a)



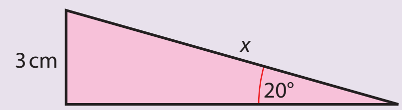
(b)



(c)



(d)



Pythagoras

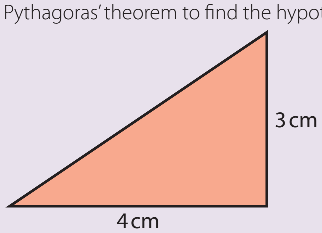
The most useful mathematical relationship for finding the resultant of two perpendicular vectors is Pythagoras' theorem.

$\text{Hypotenuse}^2 = \text{adjacent}^2 + \text{opposite}^2$

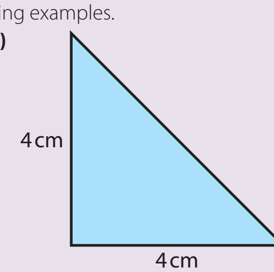
Exercise

11 Use Pythagoras' theorem to find the hypotenuse in the following examples.

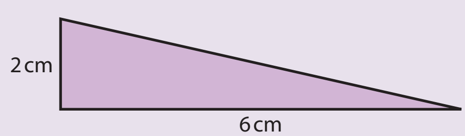
(a)



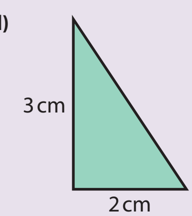
(b)



(c)



(d)



Using trigonometry to solve vector problems

Once the vectors have been arranged point to tail it is a simple matter of applying the trigonometrical relationships to the triangles that you get.

Exercises

Draw the vectors and solve the following problems using Pythagoras' theorem.

12 A boat travels 4 km west followed by 8 km north. What is the resultant displacement?

13 A plane flies 100 km north then changes course to fly 50 km east. What is the resultant displacement?

Vectors in one dimension

In this course we will often consider the simplest examples where the motion is restricted to one dimension, for example a train travelling along a straight track. In examples like this there are only two possible directions – forwards and backwards. To distinguish between the two directions,we give them different signs (forward + and backwards −). Adding vectors is now simply a matter of adding the magnitudes, with no need for complicated triangles.

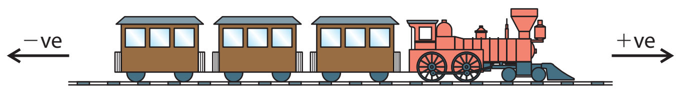


Figure 1.14 The train can only move forwards or backwards.

Which direction is +?
You can decide for yourself which direction you want to be positive but generally we follow the convention:

Right +	Left −
North +	South −
Up +	Down −

Worked example

If a train moves 100 m forwards along a straight track then 50 m back, what is its final displacement?

Solution

Figure 1.15 shows the vector diagram.

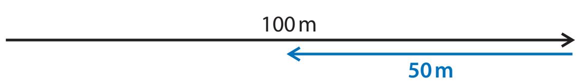


Figure 1.15 Adding vectors in one dimension.

The resultant is clearly 50 m forwards.

● **Examiner's hint:** There isn't really any need to draw vector diagrams when doing one-dimensional problems. However, you must never forget that the sign gives the direction.

Subtracting vectors

Now we know that a negative vector is simply the opposite direction to a positive vector, we can subtract vector **B** from vector **A** by changing the direction of vector **B** and adding it to **A.**

$\mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B})$

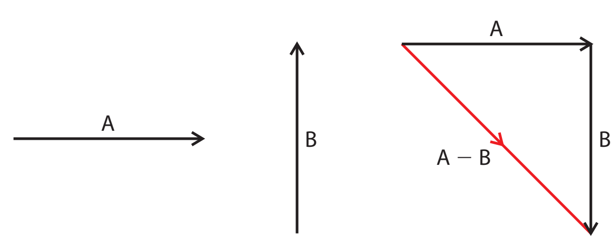


Figure 1.16 Subtracting vectors.

Taking components of a vector

Consider someone walking up the hill in Figure 1.17. They walk 5 km up the slope but want to know how high they have climbed rather than how far they have walked. To calculate this they can use trigonometry.

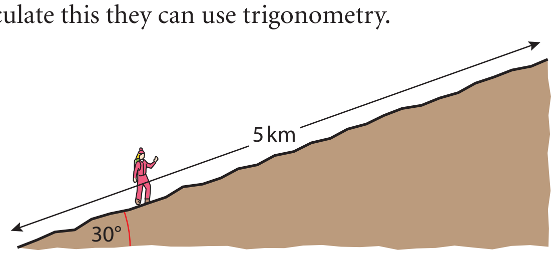


Figure 1.17 5 km up the hill but how high?

Height = 5 × sin 30°

The height is called the vertical component of the displacement.

The horizontal displacement can also be calculated.

Horizontal displacement = 5 × cos 30°

This process is called 'taking components of a vector' and is often used in solving physics problems.

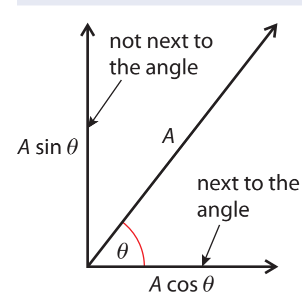


Figure 1.18 An easy way to remember which is cos is to say that it's 'becos it's next to the angle'.

Exercises

14 If a boat travels 10 km in a direction 30° to the east of north, how far north has it travelled?

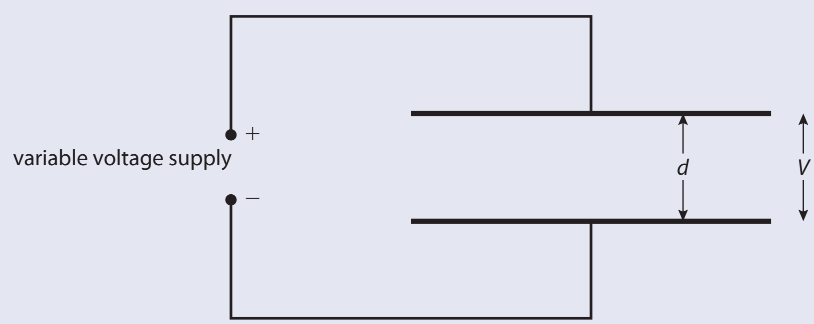
15 On his way to the South Pole, Amundsen travelled 8 km in a direction that was 20° west of south. What was his displacement south?

16 A mountaineer climbs 500 m up a slope that is inclined at an angle of 60° to the horizontal. How high has he climbed?

● **Examiner's hint:** Question 1 is about analysing data. It is a typical Paper 2 question. You don't have to know anything about permittivity to answer it.

Practice questions

1 This question is about measuring the permittivity of free space ϵ_0.
The diagram below shows two parallel conducting plates connected to a variable voltage supply. The plates are of equal areas and are a distance *d* apart.



The charge *Q* on one of the plates is measured for different values of the potential difference *V* applied between the plates. The values obtained are shown in the table below. The uncertainty in the value of *V* is not significant but the uncertainty in *Q* is ±10%.

V / V	*Q* / nC ± 10%
10.0	30
20.0	80
30.0	100
40.0	160
50.0	180

(a) Plot the data points opposite on a graph of *V* (*x*-axis) against *Q* (*y*-axis). (4)

(b) By calculating the relevant uncertainty in *Q*, add error bars to the data points (10.0, 30) and (50.0, 180). (3)

(c) On your graph, draw the line that best fits the data points and has the maximum permissible gradient. Determine the gradient of the line that you have drawn. (3)

(d) The gradient of the graph is a property of the two plates and is known as *capacitance*.
Deduce the units of capacitance. (1)

The relationship between Q and V for this arrangement is given by the expression

$$Q = \frac{\epsilon_0 A}{d} V$$

where A is the area of one of the plates.

In this particular experiment $A = 0.20 \pm 0.05\,\text{m}^2$ and $d = 0.50 \pm 0.01$ mm.

(e) Use your answer to **(c)** to determine the maximum value of ϵ_0 that this experiment yields. (4)

(Total 15 marks)

2 A student measures a distance several times. The readings lie between 49.8 cm and 50.2 cm. This measurement is best recorded as

A 49.8 ± 0.2 cm.

B 49.8 ± 0.4 cm.

C 50.0 ± 0.2 cm.

D 50.0 ± 0.4 cm. (1)

3 The time period T of oscillation of a mass m suspended from a vertical spring is given by the expression

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

where k is a constant.

Which **one** of the following plots will give rise to a straight-line graph?

A T^2 against m

B $\sqrt{T}$ against $\sqrt{m}$

C T against m

D $\sqrt{T}$ against m (1)

4 The power dissipated in a resistor of resistance R carrying a current I is equal to I^2R. The value of I has an uncertainty of $\pm 2\%$ and the value of R has an uncertainty of $\pm 10\%$. The value of the uncertainty in the calculated power dissipation is

A $\pm 8\%$.

B $\pm 12\%$.

C $\pm 14\%$.

D $\pm 20\%$. (1)

5 An ammeter has a zero offset error. This fault will affect

A neither the precision nor the accuracy of the readings.

B only the precision of the readings.

C only the accuracy of the readings.

D both the precision and the accuracy of the readings. (1)

6 When a force F of (10.0 ± 0.2) N is applied to a mass m of (2.0 ± 0.1) kg, the percentage uncertainty attached to the value of the calculated acceleration $\frac{F}{m}$ is

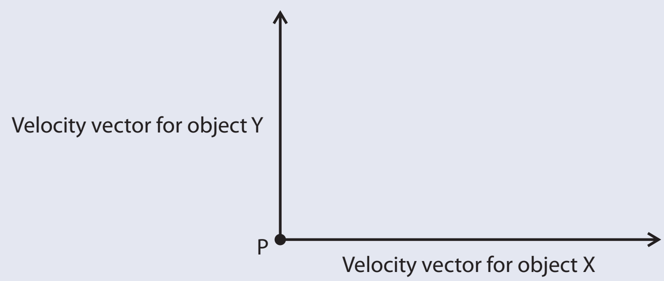
A 2%.
B 5%.
C 7%.
D 10%. (1)

7 Which of the following is the best estimate, to one significant digit, of the quantity shown below?

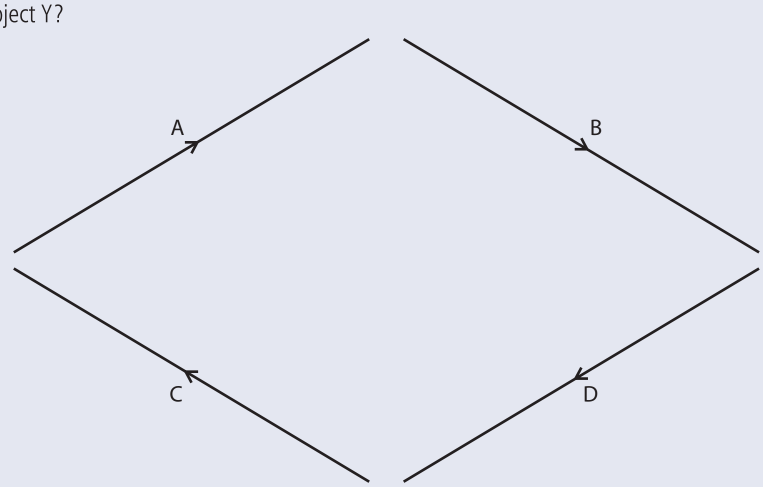
$$\frac{\pi \times 8.1}{\sqrt{(15.9)}}$$

A 1.5
B 2.0
C 5.8
D 6.0 (1)

8 Two objects X and Y are moving away from the point P. The diagram below shows the velocity vectors of the two objects.



Which of the following velocity vectors best represents the velocity of object X relative to object Y?



(1)

9 The order of magnitude of the weight of an apple is

A 10^{-4} N.
B 10^{-2} N.
C 1 N.
D 10^{2} N. (1)

2 Mechanics

2.1 Kinematics

Assessment statements

2.1.1 Define displacement, velocity, speed and acceleration.
2.1.2 Explain the difference between instantaneous and average values of speed, velocity and acceleration.
2.1.3 Outline the conditions under which the equations for uniformly accelerated motion may be applied.
2.1.9 Determine relative velocity in one and in two dimensions.

In Chapter 1, we observed that things move and now we are going to mathematically model that movement. Before we do that, we must define some quantities that we are going to use.

Displacement and distance

It is important to understand the difference between distance travelled and displacement. To explain this, consider the route marked out on the map shown in Figure 2.1

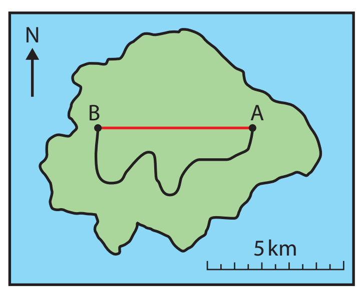


Figure 2.1

Displacement is the distance moved in a particular direction.

The unit of displacement is the metre (m).

Displacement is a vector quantity.

On the map, the displacement is the length of the straight line from A to B, a distance of 5 km west. (Note: since displacement is a vector you should always say what the direction is.)

Distance is how far you have travelled from A to B.

The unit of distance is also the metre.

Distance is a scalar quantity.

In this example, the distance travelled is the length of the path taken, which is about 10 km.

Sometimes this difference leads to a surprising result. For example, if you run all the way round a running track you will have travelled a distance of 400 m but your displacement will be 0 m.

In everyday life, it is often more important to know the distance travelled. For example, if you are going to travel from Paris to Lyon by road you will want to know that the distance by road is 450 km, not that your final displacement will be

336 km SE. However, in physics, we break everything down into its simplest units, so we start by considering motion in a straight line only. In this case it is more useful to know the displacement, since that also has information about which direction you have moved.

Velocity and speed

Both speed and velocity are a measure of how fast a body is moving, but velocity is a vector quantity and speed is a scalar.

Velocity is defined as the displacement per unit time.

$$\text{velocity} = \frac{\text{displacement}}{\text{time}}$$

The unit of velocity is $\mathrm{m\,s^{-1}}$.

Velocity is a vector quantity.

Speed is defined as the distance travelled per unit time.

$$\text{speed} = \frac{\text{distance}}{\text{time}}$$

The unit of speed is also $\mathrm{m\,s^{-1}}$.

Speed is a scalar quantity.

Exercise

1 Convert the following speeds into $\mathrm{m\,s^{-1}}$.
 (a) A car travelling at $100\,\mathrm{km\,h^{-1}}$
 (b) A runner running at $20\,\mathrm{km\,h^{-1}}$

Figure 2.2 It's not possible to take this route across Bangkok with a constant velocity.

Average velocity and instantaneous velocity

Consider travelling by car from the north of Bangkok to the south – a distance of about 16 km. If the journey takes 4 hours, you can calculate your velocity to be $\frac{16}{4} = 4\,\mathrm{km\,h^{-1}}$ in a southwards direction. This doesn't tell you anything about the journey, just the difference between the beginning and the end (unless you managed to travel at a constant speed in a straight line). The value calculated is the **average velocity** and in this example it is quite useless. If we broke the trip down into lots of small pieces, each lasting only one second, then for each second the car could be considered to be travelling in a straight line at a constant speed. For these short stages we could quote the car's **instantaneous velocity** – that's how fast it's going at that moment in time and in which direction.

The bus in the photo has a constant velocity for a very short time.

Measuring velocity

You can measure velocity with a photogate connected to a timer or computer. When a card passes through the gate it is sensed by the timer, switching it on or off.

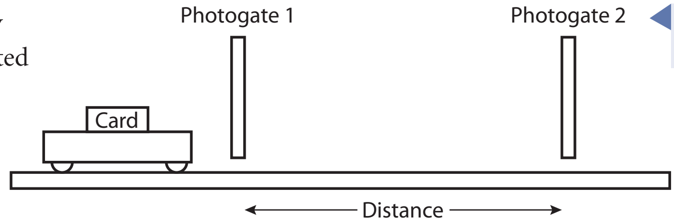


Figure 2.3 Experimental set up for measuring velocity.

$$\text{average velocity} = \frac{\text{distance}}{\text{time taken to travel between photogates}}$$

$$\text{instantaneous velocity} = \frac{\text{length of card}}{\text{time for card to pass through gate}}$$

Velocity is relative

When quoting the velocity of a body, it is important to say what the velocity is measured relative to. Consider the people in Figure 2.4

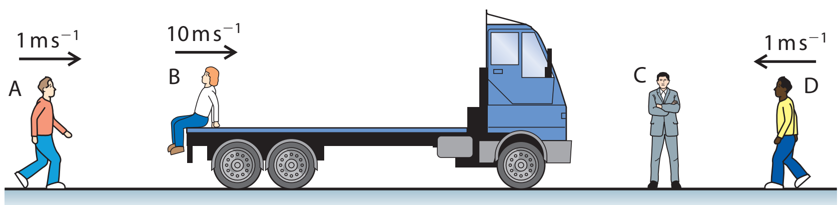


Figure 2.4 Two observers measuring the same velocity.

C measures the velocity of *A* to be 1 m s^{-1} but to B (moving on the truck towards *C*) the velocity of *A* is -9 m s^{-1} (*B* will see *A* moving away in a negative direction). You might think that *A* can't have two velocities, but he can – velocity is relative. In this example there are two observers, *B* and *C*. Each observer has a different 'frame of reference'. To convert a velocity, to *B*'s frame of reference, we must subtract the velocity of *B* relative to *C*; this is 10 m s^{-1}.

● **Examiner's hint:** Velocity is a vector. If the motion is in one dimension, the direction of velocity is given by its sign. Generally, right is positive and left negative.

So the velocity of *A* relative to $B = 1 - 10 = -9$ m s^{-1}

We can try the same with *D* who has a velocity of -1 m s^{-1} measured by *C* and $-1 - 10 = -11$ m s^{-1} measured by *B*.

This also works in two dimensions as follows:

A now walks across the road as illustrated by the aerial view in Figure 2.5. The velocity of *A* relative to *C* is 1 m s^{-1} north.

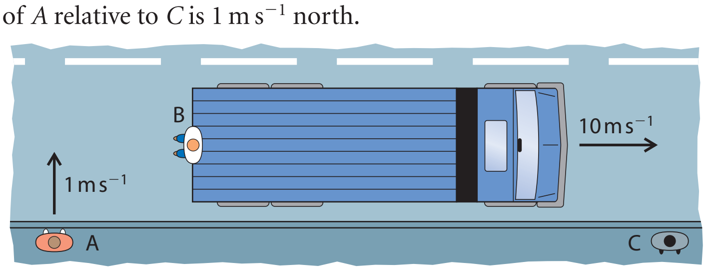


Figure 2.5 *A* now walks across the road..

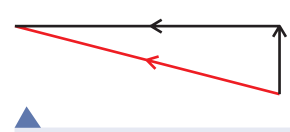


Figure 2.6 Subtracting vectors gives the relative velocity.

The velocity of *A* relative to *B* can now be found by subtracting the vectors as shown in Figure 2.6.

Exercise

2 An observer standing on a road watches a bird flying east at a velocity of 10 m s^{-1}. A second observer, driving a car along the road northwards at 20 m s^{-1} sees the bird. What is the velocity of the bird relative to the driver?

Acceleration

In everyday usage, the word *accelerate* means to go faster. However in physics: acceleration is defined as the rate of change of velocity.

$$\text{acceleration} = \frac{\text{change of velocity}}{\text{time}}$$

The unit of acceleration is $\mathrm{m\,s^{-2}}$.

Acceleration is a vector quantity.

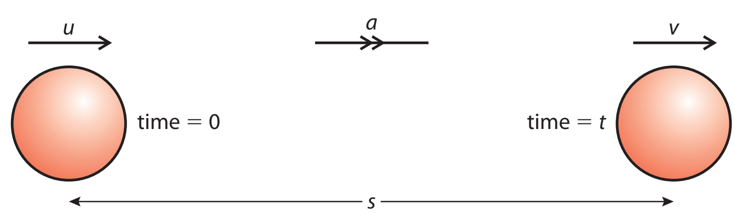
This means that whenever a body changes its velocity it accelerates. This could be because it is getting faster, slower or just changing direction. In the example of the journey across Bangkok, the car would have been slowing down, speeding up and going round corners almost the whole time, so it would have had many different accelerations. However, this example is far too complicated for us to consider in this course (and probably any physics course). For most of this chapter we will only consider the simplest example of accelerated motion, constant acceleration.

Bodies
When we refer to a body in physics we generally mean a ball not a human body.

Constant acceleration in one dimension

In one-dimensional motion, the acceleration, velocity and displacement are all in the same direction. This means they can simply be added without having to draw triangles. Figure 2.7 shows a body that is starting from an initial velocity u and accelerating at a constant rate to velocity v in t seconds. The distance travelled in this time is s. Since the motion is in a straight line, this is also the displacement.

Figure 2.7 A red ball is accelerated at a constant rate.



Using the definitions already stated, we can write equations related to this example.

Average velocity

From the definition, the average velocity $= \dfrac{\text{displacement}}{\text{time}}$

So average velocity $= \frac{s}{t}$ (1)

Since the velocity changes at a constant rate from the beginning to the end, we can also calculate the average velocity by adding the velocities and dividing by two.

Average velocity $= \dfrac{(u + v)}{2}$ (2)

Acceleration

Acceleration is defined as the rate of change of velocity.

So $a = \dfrac{(v - u)}{t}$ (3)

We can use these equations to solve any problem involving constant acceleration. However, to make problem solving easier, we can derive two more equations by substituting from one into the other.

Equating equations (1) and(2)

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

so $s = \dfrac{(u+v)\,t}{2}$ (4)

Rearranging (3) gives $v = u + at$

If we substitute for v in equation (4) we get $s = ut + \frac{1}{2}at^2$ (5)

Rearranging (3) again gives $t = \dfrac{(v-u)}{a}$

If t is now substituted in equation (4) we get $v^2 = u^2 + 2as$ (6)

These equations are sometimes known as the *suvat* equations. If you know any 3 of *suva* and *t* you can find either of the other two in one step.

suvat equations

$a = \dfrac{(v-u)}{t}$

$s = \dfrac{(u+v)\,t}{2}$

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

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

Worked example

1 A car travelling at $10\,\mathrm{m\,s^{-1}}$ accelerates at $2\,\mathrm{m\,s^{-2}}$ for 5 s. What is its displacement?

Solution

The first thing to do is draw a simple diagram like Figure 2.8.

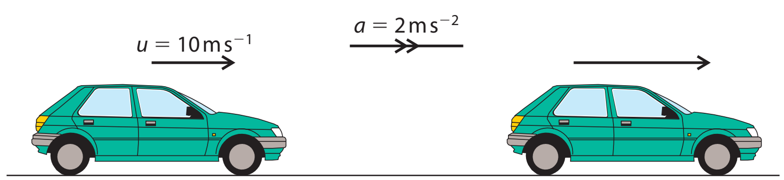


Figure 2.8 A simple diagram is always the best start.

This enables you to see what is happening at a glance rather than reading the text. The next stage is to make a list of *suvat*.

$$s = ?$$
$$u = 10\,\mathrm{m\,s^{-1}}$$
$$v = ?$$
$$a = 2\,\mathrm{m\,s^{-2}}$$
$$t = 5\,\mathrm{s}$$

To find s you need an equation that contains *suat*. The only equation with all 4 of these quantities is $s = ut + \frac{1}{2}at^2$

Using this equation gives:

$$s = 10 \times 5 + \tfrac{1}{2} \times 2 \times 5^2$$
$$s = 75\,\mathrm{m}$$

● **Examiner's hint:** You don't need to include units in all stages of a calculation, just the answer.

The sign of displacement, velocity and acceleration

We must not forget that displacement, velocity and acceleration are vectors. This means that they have direction. However, since this is a one-dimensional example, there are only two possible directions, forward and backward. We know which direction the quantity is in from the sign.

A positive displacement means that the body has moved right.

A positive velocity means the body is moving to the right.

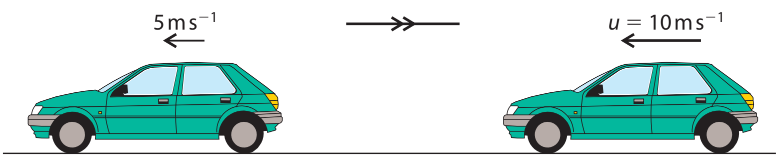


Figure 2.9

A positive acceleration means that the body is either moving to the right and getting faster or moving to the left and getting slower. This can be confusing, so consider the following example.

The car is travelling in a negative direction so the velocities are negative.

$$u = -10\ \text{m s}^{-1}$$
$$v = -5\ \text{m s}^{-1}$$
$$t = 5\ \text{s}$$

The acceleration is therefore given by

$$a = \frac{(v - u)}{t} = \frac{-5 - -10}{5} = 1\ \text{m s}^{-2}$$

The positive sign tells us that the acceleration is in a positive direction (right) even though the car is travelling in a negative direction (left).

Example

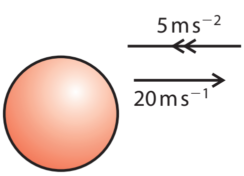


Figure 2.10 The acceleration is negative so pointing to the left.

A body with a constant acceleration of $-5\ \text{m s}^{-2}$ is travelling to the right with a velocity of $20\ \text{m s}^{-1}$. What will its displacement be after 20 s?

$$s = ?$$
$$u = 20\ \text{m s}^{-1}$$
$$v = ?$$
$$a = -5\ \text{m s}^{-2}$$
$$t = 20\ \text{s}$$

To calculate s we can use the equation $s = ut + \frac{1}{2}at^2$

$$s = 20 \times 20 + \tfrac{1}{2}(-5) \times 20^2 = 400 - 1000 = -600\ \text{m}$$

This means that the final displacement of the body is to the left of the starting point. It has gone forward, stopped and then gone backwards.

g

The acceleration due to gravity is not constant all over the Earth. $9.81\ \text{m s}^{-2}$ is the average value. The acceleration also gets smaller the higher you go. However we ignore this change when conducting experiments in the lab since labs aren't that high.

To make the examples easier to follow, $g = 10\ \text{m s}^{-2}$ is used throughout; you should only use this approximate value in exam questions if told to do so.

Exercises

3 Calculate the final velocity of a body that starts from rest and accelerates at $5\ \text{m s}^{-2}$ for a distance of 100 m.

4 A body starts with a velocity of $20\ \text{m s}^{-1}$ and accelerates for 200 m with an acceleration of $5\ \text{m s}^{-2}$. What is the final velocity of the body?

5 A body accelerates at $10\ \text{m s}^{-2}$ reaching a final velocity of $20\ \text{m s}^{-1}$ in 5 s. What was the initial velocity of the body?

2.2 Free fall motion

Assessment statements

2.1.4 Identify the acceleration of a body falling in a vacuum near the Earth's surface with the acceleration g of free fall.

2.1.5 Solve problems involving the equations of uniformly accelerated motion.

2.1.6 Describe the effects of air resistance on falling objects.

Although a car was used in one of the previous illustrations, the acceleration of a car is not usually constant, so we shouldn't use the *suvat* equations. The only example of constant acceleration that we see in everyday life is when a body is dropped. Even then the acceleration is only constant for a short distance.

Acceleration of free fall

When a body is allowed to fall freely we say it is in free fall. Bodies falling freely on the Earth fall with an acceleration of about $9.81\ \mathrm{m\,s^{-2}}$. (It depends where you are.) The body falls because of gravity. For that reason we use the letter g to denote this acceleration. Since the acceleration is constant, we can use the *suvat* equations to solve problems.

The effect of air resistance
If you jump out of a plane (with a parachute on) you will feel the push of the air as it rushes past you. As you fall faster and faster, the air will push upwards more and more until you can't go any faster. At this point you have reached terminal velocity. We will come back to this example after introducing forces.

Exercises

In these calculations use $g = 10\ \mathrm{m\,s^{-2}}$.

6 A ball is thrown upwards with a velocity of $30\ \mathrm{m\,s^{-1}}$. What is the displacement of the ball after 2 s?

7 A ball is dropped. What will its velocity be after falling 65 cm?

8 A ball is thrown upwards with a velocity of $20\ \mathrm{m\,s^{-1}}$. After how many seconds will the ball return to its starting point?

Measuring *g*

Measuring g by timing a ball falling from different heights is a common physics experiment that you could well perform in the practical programme of the IB course. There are various different ways of doing this but a common method is to use a timer that starts when the ball is released and stops when it hits a platform. An example of this apparatus is shown in the photo. The distance travelled by the ball and the time taken are related by the *suvat* equation $s = ut + \frac{1}{2}at^2$. This simplifies to $s = \frac{1}{2}at^2$ since the initial velocity is zero. This means that s is proportional to t^2 so if you plot a graph of s against t^2 you will get a straight line whose gradient is $\frac{1}{2}g$.

Free fall apparatus.

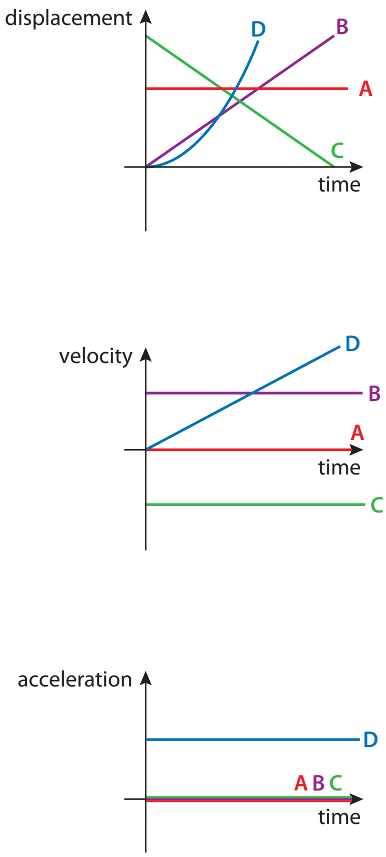
2.3 Graphical representation of motion

Assessment statements

2.1.7 Draw and analyse distance–time graphs, displacement–time graphs, velocity–time graphs and acceleration–time graphs.

2.1.8 Calculate and interpret the gradients of displacement–time graphs and velocity–time graphs, and the areas under velocity–time graphs and acceleration–time graphs.

Figure 2.11 Graphical representation of motion.



Graphs are used in physics to give a visual representation of relationships. In kinematics they can be used to show how displacement, velocity and acceleration change with time. Figure 2.11 shows the graphs for four different examples of motion. They are placed vertically since they all have the same time axis.

Line A

A body that is not moving.
Displacement is always the same.
Velocity is zero.
Acceleration is zero.

Line B

A body that is travelling with a constant positive velocity.
Displacement increases linearly with time.
Velocity is a constant positive value.
Acceleration is zero.

Line C

A body that has a constant negative velocity.
Displacement is decreasing linearly with time.
Velocity is a constant negative value.
Acceleration is zero.

Line D

A body that is accelerating with constant acceleration.
Displacement is increasing at a non-linear rate. The shape of this line is a parabola since displacement is proportional to t^2 ($s = ut + \frac{1}{2}at^2$).
Velocity is increasing linearly with time.
Acceleration is a constant positive value.

Examiner's hint: You need to be able to
- figure out what kind of motion a body has by looking at the graphs
- sketch graphs for a given motion.

The best way to go about sketching graphs is to split the motion into sections then plot where the body is at different times; joining these points will give the displacement–time graph. Once you have done that you can work out the v–t and a–t graphs by looking at the s–t graph rather than the motion.

Gradient of displacement–time

The gradient of a graph is $\dfrac{\text{change in } y}{\text{change in } x}$

$$= \frac{\Delta y}{\Delta x}$$

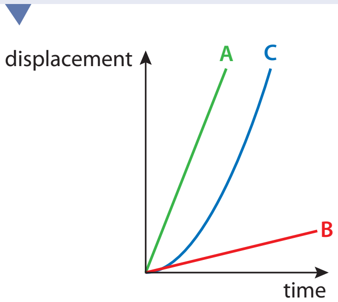
In the case of the displacement–time graph this will give

$$\text{gradient} = \frac{\Delta s}{\Delta t}$$

This is the same as velocity.

So the gradient of the displacement–time graph equals the velocity. Using this information, we can see that line A in Figure 2.12 represents a body with greater velocity than line B and that since the gradient of line C is increasing, this must be the graph for an accelerating body.

Figure 2.12



Instantaneous velocity

When a body accelerates its velocity is constantly changing. The displacement–time graph for this motion is therefore a curve. To find the instantaneous velocity from the graph we can draw a tangent to the curve and find the gradient of the tangent as shown in Figure 2.13.

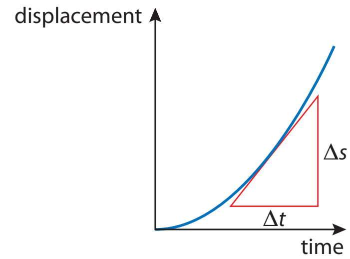


Figure 2.13

Area under velocity–time graph

The area under the velocity–time graph for the body travelling at constant velocity v shown in Figure 2.14 is given by

$$\text{area} = v\Delta t$$

But we know from the definition of velocity that $v = \dfrac{\Delta s}{\Delta t}$

Rearranging gives $\Delta s = v\Delta t$ so the area under a velocity–time graph gives the displacement.

This is true not only for simple cases such as this but for all examples.

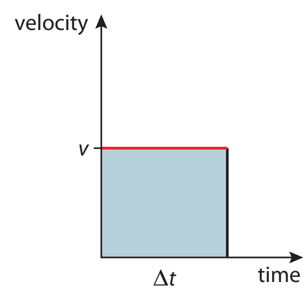


Figure 2.14

Gradient of velocity–time graph

The gradient of the velocity–time graph is given by $\dfrac{\Delta v}{\Delta t}$. This is the same as acceleration.

Area under acceleration–time graph

The area under an acceleration–time graph in Figure 2.15 is given by $a\Delta t$. But we know from the definition of acceleration that $a = \dfrac{(v - u)}{t}$

Rearranging this gives $v - u = a\Delta t$ so the area under the graph gives the change in velocity.

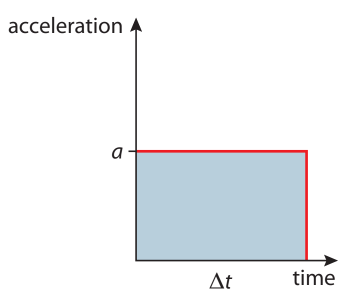


Figure 2.15

If you have covered calculus in your maths course you may recognise these equations:

$$v = \frac{ds}{dt} a = \frac{dv}{dt} = \frac{d^2s}{dt^2} \text{ and } s = \int v dt,\ v = \int a dt$$

Exercises

9 Sketch a velocity–time graph for a body starting from rest and accelerating at a constant rate to a final velocity of $25\ \mathrm{m\,s^{-1}}$ in 10 seconds. Use the graph to find the distance travelled and the acceleration of the body.

10 Describe the motion of the body whose velocity–time graph is shown in Figure 2.16. What is the final displacement of the body?

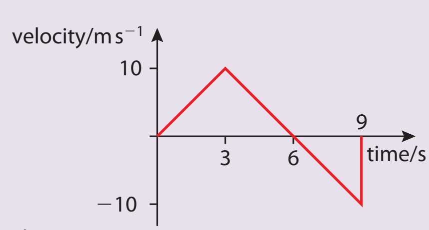


Figure 2.16

Example 1: the *suvat* example

As an example let us consider the motion we looked at when deriving the *suvat* equations.

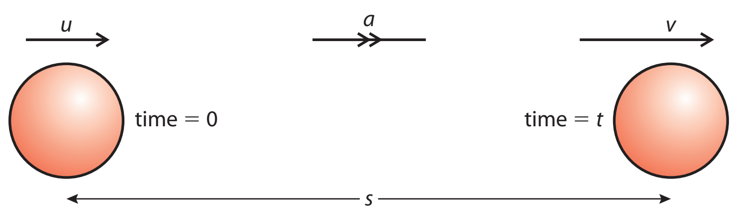


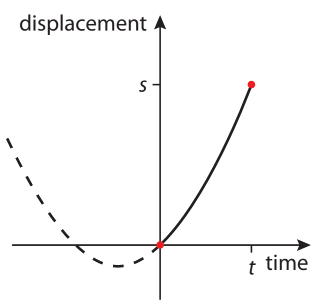
Figure 2.17 A body with constant acceleration.

Negative time
Negative time doesn't mean going back in time – it means the time before you started the clock.

Displacement–time

The body starts with velocity u and travels to the right with constant acceleration, a for a time t. If we take the starting point to be zero displacement, then the displacement–time graph starts from zero and rises to s in t seconds. We can therefore plot the two points shown in Figure 2.18. The body is accelerating so the line joining these points is a parabola. The whole parabola has been drawn to show what it would look like – the reason it is offset is because the body is not starting from rest. The part of the curve to the left of the origin tells us what the particle was doing before we started the clock.

Figure 2.18



Velocity–time

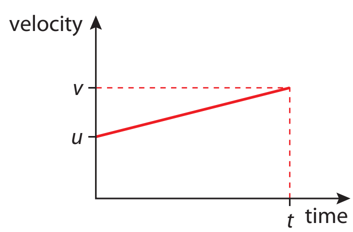
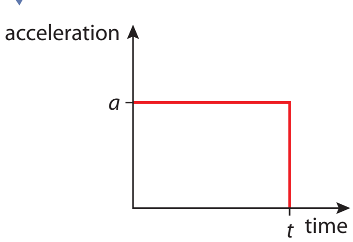


Figure 2.19

Figure 2.19 is a straight line with a positive gradient showing that the acceleration is constant. The line doesn't start from the origin since the initial velocity is u. The gradient of this line is $\frac{(v-u)}{t}$ which we know from the *suvat* equations is acceleration.

The area under the line makes the shape of a trapezium. The area of this trapezium is $\frac{1}{2}(v+u)t$. This is the *suvat* equation for s.

Figure 2.20



Acceleration–time

The acceleration is constant so the acceleration–time graph is simply a horizontal line as shown in Figure 2.20. The area under this line is $a \times t$ which we know from the *suvat* equations equals $(v-u)$.

Example 2: The bouncing ball

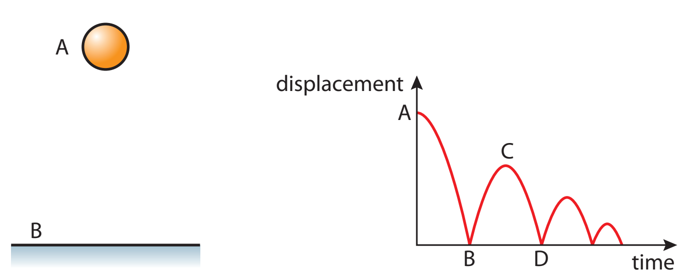


Figure 2.21

Consider a rubber ball dropped from some position above the ground A onto a hard surface B. The ball bounces up and down several times. Figure 2.21 shows the displacement–time graph for 4 bounces. From the graph we see that the ball starts above the ground then falls with increasing velocity (as deduced by the increasing negative gradient). When the ball bounces at B the velocity suddenly changes from negative to positive as the ball begins to travel back up. As the ball goes up, its velocity gets less until it stops at C and begins to fall again.

To view a simulation that enables you to plot the graphs as you watch the motion, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 2.1.

Exercise

11 By considering the gradient of the displacement–time graph in Figure 2.21 plot the velocity–time graph for the motion of the bouncing ball.

Example 3: A ball falling with air resistance

Figure 2.22 represents the motion of a ball that is dropped several hundred metres through the air. It starts from rest and accelerates for some time. As the ball accelerates, the air resistance gets bigger, which prevents the ball from getting any faster. At this point the ball continues with constant velocity.

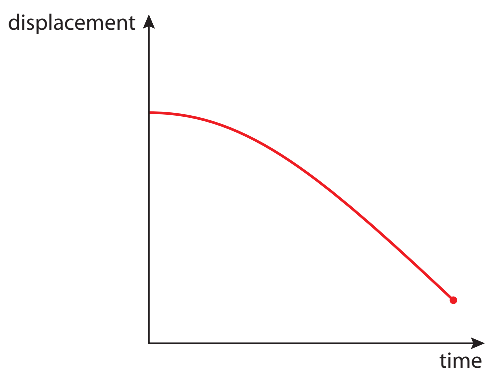


Figure 2.22

Exercise

12 By considering the gradient of the displacement–time graph plot the velocity–time graph for the motion of the falling ball.

2.4 Projectile motion

Assessment statements

9.1.1 State the independence of the vertical and the horizontal components of velocity for a projectile in a uniform field.

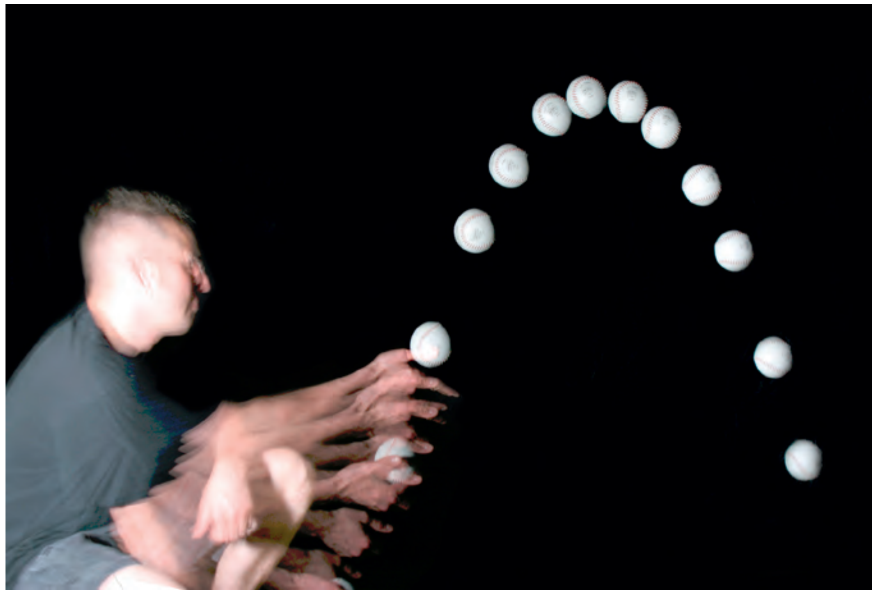
9.1.2 Describe and sketch the trajectory of projectile motion as parabolic in the absence of air resistance.

9.1.3 Describe qualitatively the effect of air resistance on the trajectory of a projectile.

9.1.4 Solve problems on projectile motion.

We all know what happens when a ball is thrown; it follows a curved path like the one in the photo below. We can see from this photo that the path is parabolic, and later we will show why that is the case.

A stroboscopic photograph of a projected ball.

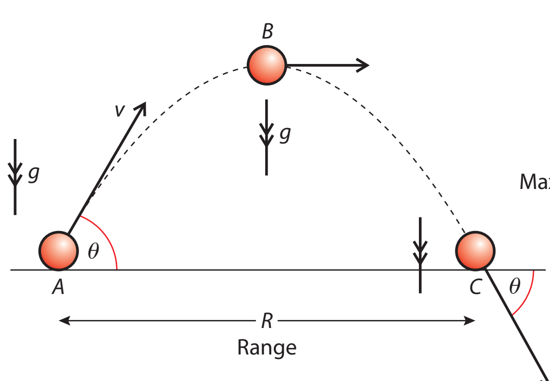


Modelling projectile motion

All examples of motion up to this point have been in one dimension but projectile motion is two-dimensional. However, if we take components of all the vectors vertically and horizontally, we can simplify this into two simultaneous one-dimensional problems. The important thing to realise is that the vertical and horizontal components are independent of each other; you can test this by dropping a stone off a cliff and throwing one forward at the same time, they both hit the bottom together. The downward motion is not altered by the fact that one is also moving forward.

Consider a ball that is projected at an angle θ to the horizontal, as shown in Figure 2.23. We can split the motion into three parts, beginning, middle and end, and analyse the vectors representing displacement, velocity and time at each stage. Note that since the path is symmetrical, the motion on the way down is the same as the way up.

Figure 2.23 A projectile launched at an angle θ.



Horizontal components

At A (time = 0)	At B (time = $\frac{t}{2}$)	At C (time = t)
Displacement = zero Velocity = $v \cos \theta$ Acceleration = 0	Displacement = $\frac{R}{2}$ Velocity = $v \cos \theta$ Acceleration = 0	Displacement = R Velocity = $v \cos \theta$ Acceleration = 0

Vertical components

At A	At B	At C
Displacement = zero Velocity = $v \sin \theta$ Acceleration = $-g$	Displacement = h Velocity = zero Acceleration = $-g$	Displacement = zero Velocity = $-v \sin \theta$ Acceleration = $-g$

We can see that the vertical motion is constant acceleration and the horizontal motion is constant velocity. We can therefore use the *suvat* equations.

suvat for horizontal motion

Since acceleration is zero there is only one equation needed to define the motion

suvat	A to C
Velocity = $v = \frac{s}{t}$	$R = v \cos \theta \, t$

suvat for vertical motion

When considering the vertical motion it is worth splitting the motion into two parts.

suvat	At B	At C
$s = \frac{1}{2}(u + v)t$	$h = \frac{1}{2}(v \sin \theta)\frac{t}{2}$	$0 = \frac{1}{2}(v \sin \theta - v \sin \theta)t$
$v^2 = u^2 + 2as$	$0 = v^2 \sin^2\theta - 2gh$	$(-v \sin \theta)^2 = (v \sin \theta)^2 - 0$
$s = ut + \frac{1}{2}at^2$	$h = v \sin \theta \, t - \frac{1}{2}g\left(\frac{t}{2}\right)^2$	$0 = v \sin \theta \, t - \frac{1}{2}gt^2$
$a = \frac{v - u}{t}$	$g = \frac{v \sin \theta - 0}{\frac{t}{2}}$	$g = \frac{v \sin \theta - -v \sin \theta}{t}$

Some of these equations are not very useful since they simply state that 0 = 0. However we do end up with three useful ones (highlighted):

$R = v \cos \theta \, t$ (1)

$0 = v^2 \sin^2\theta - 2gh$ or $h = \frac{v^2 \sin^2 \theta}{2g}$ (2)

$0 = v \sin \theta \, t - \frac{1}{2}gt^2$ or $t = \frac{2v \sin \theta}{g}$ (3)

Solving problems

In a typical problem you will be given the magnitude and direction of the initial velocity and asked to find either the maximum height or range. To calculate h you can use equation (2) but to calculate R you need to find the time of flight so must use (3) first (you could also substitute for t into equation (1) to give a fourth equation but maybe we have enough equations already).

You do not have to remember a lot of equations to solve a projectile problem. If you understand how to apply the *suvat* equations to the two components of the projectile motion, you only have to remember the *suvat* equations (and they are in the databook).

Parabolic path
Since the horizontal displacement is proportional to t the path has the same shape as a graph of vertical displacement plotted against time. This is parabolic since the vertical displacement is proportional to t^2.

Maximum range
For a given value of v the maximum range is when $v \cos \theta \, t$ is a maximum value. Now $t = \frac{2v \sin \theta}{g}$. If we substitute this for t we get

$$R = \frac{2v^2 \cos \theta \sin \theta}{g}.$$

This is a maximum when $\cos\theta \sin\theta$ is maximum, which is when $\theta = 45°$.

To view a simulation of projectile motion, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 2.2.

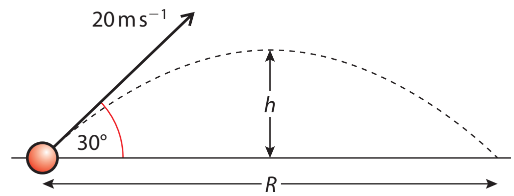
Worked example

1 A ball is thrown at an angle of 30° to the horizontal at a speed of $20\ \mathrm{m\,s^{-1}}$. Calculate its range and the maximum height reached.

2 A ball is thrown horizontally from a cliff top with a horizontal speed of $10\ \mathrm{m\,s^{-1}}$. If the cliff is 20 m high what is the range of the ball?

Solution

1 First, as always, draw a diagram, including labels defining all the quantities known and unknown.



Now we need to find the time of flight. If we apply $s = ut + \frac{1}{2}at^2$ to the whole flight we get

$$t = \frac{2v\sin\theta}{g} = \frac{(2 \times 20 \times \sin 30°)}{10} = 2\ \mathrm{s}$$

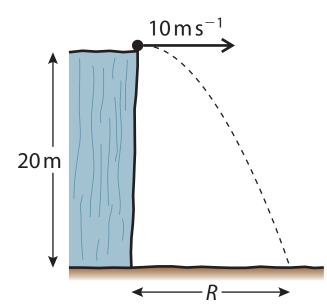
We can now apply $s = vt$ to the whole flight to find the range:

$$R = v\cos\theta t = 20 \times \cos 30° \times 2 = \mathbf{34.6\ m}$$

Finally to find the height, we use $s = ut + \frac{1}{2}at^2$ to the vertical motion, but remember, this is only half the complete flight so the time is 1 s.

$$h = v\sin\theta t - \tfrac{1}{2}gt^2 = 20 \times \sin 30° \times 1 - \tfrac{1}{2} \times 10 \times 1^2 = 10 - 5 = \mathbf{5\ m}$$

2 This is an easy one since there aren't any angles to deal with. The initial vertical component of the velocity is zero and the horizontal component is $10\ \mathrm{m\,s^{-1}}$. To calculate the time of flight we apply $s = ut + \frac{1}{2}at^2$ to the vertical component. Knowing that the final displacement is -20 m this gives



$$-20\ \mathrm{m} = 0 - \tfrac{1}{2}gt^2 \text{ so } t = \sqrt{\frac{(2 \times 20)}{10}} = 2\ \mathrm{s}$$

We can now use this value to find the range by applying the formula $s = vt$ to the horizontal component: $R = 10 \times 2 = \mathbf{20\ m}$

If you have ever played golf you will know it is not true that the maximum range is achieved with an angle of 45°, it's actually much less. This is because the ball is held up by the air like an aeroplane is. In this photo Alan Shepard is playing golf on the moon. Here the maximum range will be at 45°.

Exercises

13 Calculate the range of a projectile thrown at an angle of 60° to the horizontal with velocity $30\ \mathrm{m\,s^{-1}}$.

14 You throw a ball at a speed of $20\ \mathrm{m\,s^{-1}}$.

(a) At what angle must you throw it so that it will just get over a wall that is 5 m high?

(b) How far away from the wall must you be standing?

15 A gun is aimed so that it points directly at the centre of a target 200 m away. If the bullet travels at $200\ \mathrm{m\,s^{-1}}$ how far below the centre of the target will the bullet hit?

16 If you can throw a ball at $20\ \mathrm{m\,s^{-1}}$ what is the maximum distance you can throw it?

Projectile motion with air resistance

In all the examples above we have ignored the fact that the air will resist the motion of the ball. The actual path of a ball including air resistance is likely to be as shown in Figure 2.24.

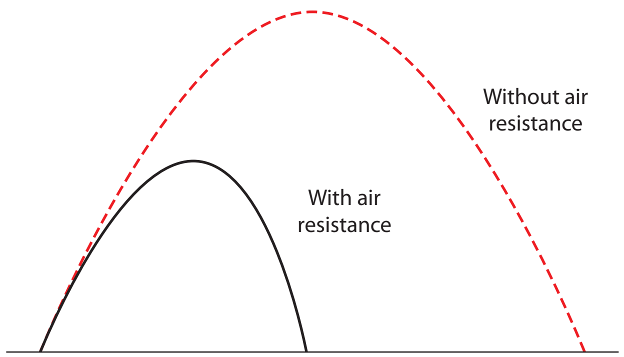


Figure 2.24

Notice both the height and range are less. It is also no longer a parabola – the way down is steeper than the way up.

2.5 Forces and dynamics

Assessment statements

2.2.1 Calculate the weight of a body using the expression $W = mg$.

2.2.2 Identify the forces acting on an object and draw free body diagrams representing the forces acting.

2.2.3 Determine the resultant force in different situations.

Forces

From experience, we know that things don't seem to move unless we push them, so movement is related to pushing. In this next section we will investigate this relationship.

The size of one newton

If you hold an object of mass 100 g in your hand then you will be exerting an upward force of about one newton (1 N).

What is a force?

A force is simply a push or a pull.

The unit of force is the newton (N).

Force is a vector quantity.

You might believe that there are hundreds of different ways to push or pull an object but there are actually surprisingly few.

1. Tension

If you attach a rope to a body and pull it, the rope is in tension. This is also the name of the force exerted on the body.

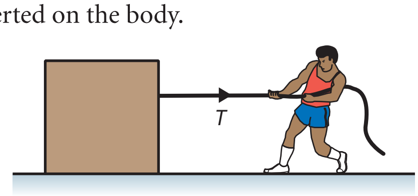


Figure 2.25 The force experienced by the block is tension, *T*.

2. Normal force

Whenever two surfaces are in contact, there will be a force between them (if not then they are not in contact). This force acts at right angles to the surface so is called the *normal force*.

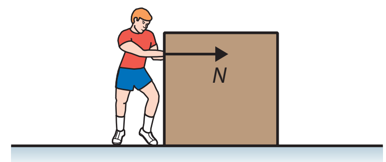


Figure 2.26 The man pushes the block with his hands. The force is called the normal force, *N*.

3. Gravitational force

We know that all objects experience a force that pulls them downwards; we call this force the *weight*. The direction of this force is always towards the centre of the Earth. The weight of a body is directly proportional to the mass of the body. $W = mg$ where $g =$ the acceleration of free fall.

You will discover why this is the case later in the chapter.

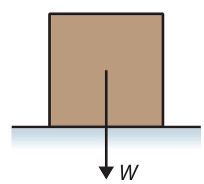


Figure 2.27 This box is pulled downwards by gravity. We call this force the weight, *W*.

Where to draw forces
It is important that you draw the point of application of the forces in the correct place. Notice where the forces are applied in these diagrams.

4. Friction force

Whenever two touching surfaces move, or attempt to move, relative to each other, there is a force that opposes the motion. This is called *frictional force*. The size of this force is dependent on the material of the surfaces and how much force is used to push them together.

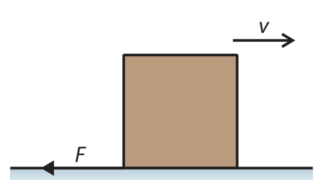


Figure 2.28 This box sliding along the floor will slow down due to the friction, *F*, between it and the floor.

5. Upthrust

Upthrust is the name of the force experienced by a body immersed in a fluid (gas or liquid). This is the force that pushes up on a boat enabling it to float in water. The size of this force is equal to the weight of fluid displaced by the boat.

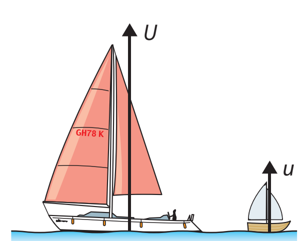
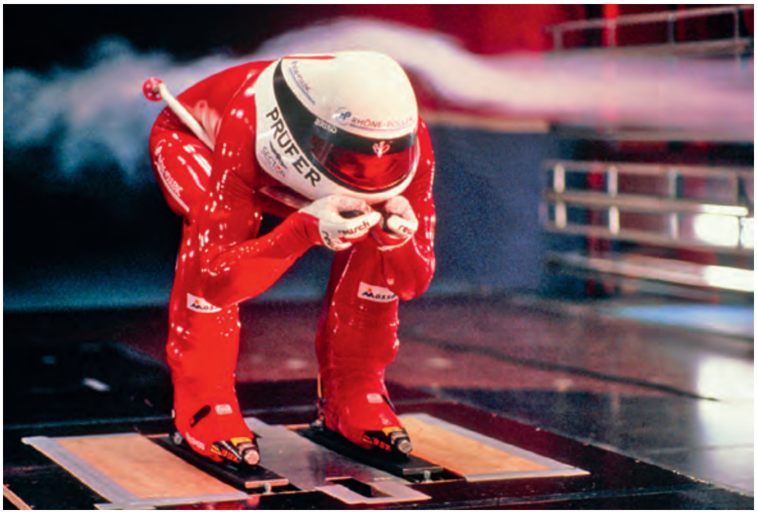


Figure 2.29 Upthrust *U* depends on how much water is displaced.

6. Air resistance

Air resistance is the force that opposes the motion of bodies through the air. This force is dependent on the speed, size and shape of the body.

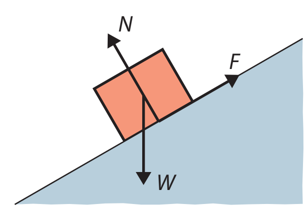


Speed skiers wear special clothes and squat down like this to reduce air resistance.

Free body diagrams

Problems often involve more than one body and more than one force. To keep things simple we always draw each body separately and only the forces acting on that body, not the forces that body exerts on something else. This is called a *free body diagram*.

Figure 2.30



A good example of this is a block resting on a ramp as shown in Figure 2.30. The block will also exert a force on the slope but this is not shown, since it is a free body diagram of the block not the ramp. Another common example that we will come across many times is a mass swinging on the end of a rope, as shown in Figure 2.31.

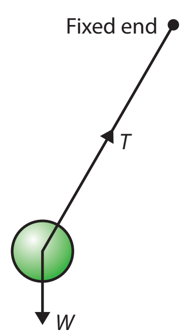


Figure 2.31

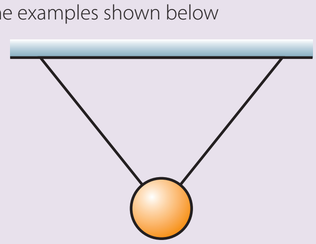
Exercise

17 Draw free body diagrams for

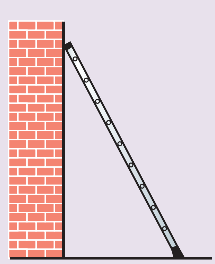
(a) a box resting on the floor

(b) the examples shown below

i



ii



(c) a free fall parachutist falling through the air

(d) a boat floating in water.

Adding forces

Force is a vector quantity, so if two forces act on the same body you must add them vectorially as with displacements and velocities.

Examples

1 A body is pulled in two opposing directions by two ropes as shown in Figure 2.32. The resultant force acting is the vector sum of the forces.

The sum is found by arranging the vectors point to tail. This gives a resultant of 2 N to the left.

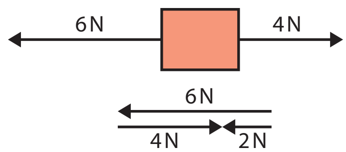


Figure 2.32

2 If a body is pulled by two perpendicular ropes as in Figure 2.33, then the vector addition gives a triangle that can be solved by Pythagoras.

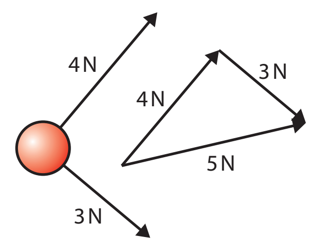
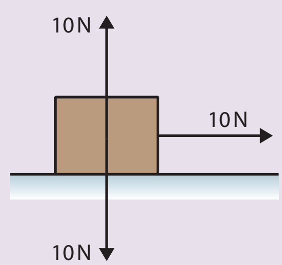


Figure 2.33

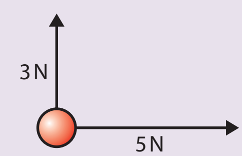
Exercise

18 Find the resultant force in the following examples:

(a)



(b)



Balanced forces

If the resultant force on a body is zero, the forces are said to be balanced. For example, if we add together the vectors representing the forces on the box in Figure 2.34 then we can see that they add up to zero. The forces are therefore balanced.

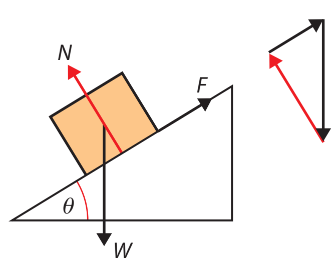


Figure 2.34

This can lead to some complicated triangles so it is easier to take components of the forces; if the components in any two perpendicular directions are balanced, then the forces are balanced. Figure 2.35 shows how this would be applied to the same example. To make things clear, the vectors have been drawn away from the box.

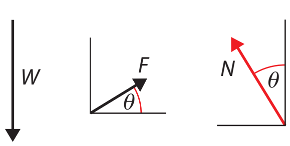


Figure 2.35

Vertical components add up to zero.

$F\sin\theta + N\cos\theta - W = 0$

$W = F\sin\theta + N\cos\theta$

Horizontal components add up to zero.

$0 + F\cos\theta - N\sin\theta = 0$

$F\cos\theta = N\sin\theta$

So we can see that

forces up = forces down
forces left = forces right

Exercises

19 A ball of weight 10 N is suspended on a string and pulled to one side by another horizontal string as shown in Figure 2.36. If the forces are balanced:

(a) write an equation for the horizontal components of the forces acting on the ball

(b) write an equation for the vertical components of the forces acting on the ball

(c) use the second equation to calculate the tension in the upper string, *T*

(d) use your answer to **(c)** plus the first equation to find the horizontal force *F*.

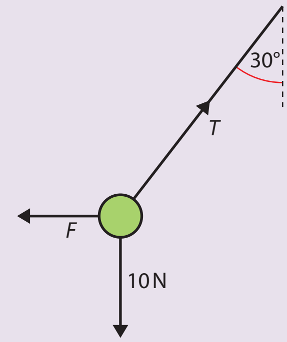


Figure 2.36

20 The condition for the forces to be balanced is that the sum of components of the forces in any two perpendicular components is zero. In the 'box on a ramp' example the vertical and horizontal components were taken. However, it is sometimes more convenient to consider components parallel and perpendicular to the ramp.

Consider the situation in Figure 2.37. If the forces on this box are balanced:

(a) write an equation for the components of the forces parallel to the ramp

(b) write an equation for the forces perpendicular to the ramp

(c) use your answers to find the friction (*F*) and normal force (*N*).

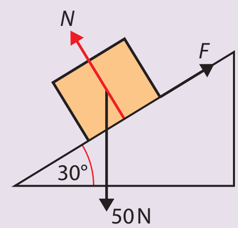


Figure 2.37

21 A rock climber is hanging from a rope attached to the cliff by two bolts as shown in Figure 2.38. If the forces are balanced

(a) write an equation for the vertical component of the forces on the knot

(b) write an equation for the horizontal forces exerted on the knot

(c) calculate the tension *T* in the ropes joined to the bolts.

The result of this calculation shows why ropes should not be connected in this way.

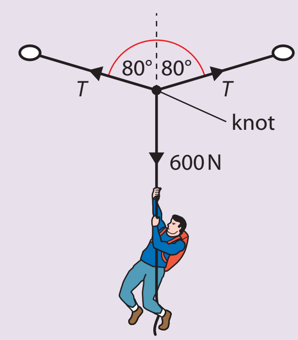


Figure 2.38

2.6 Newton's laws of motion

Assessment statements

2.2.4 State Newton's first law of motion.

2.2.5 Describe examples of Newton's first law.

2.2.6 State the condition for translational equilibrium.

2.2.7 Solve problems involving translational equilibrium.

We now have the quantities to enable us to model motion and we have observed that to make something start moving we have to exert a force – but we haven't connected the two. Newton's laws of motion connect motion with its cause. In this course there are certain fundamental concepts that everything else rests upon, Newton's laws of motion are among the most important of these.

Using laws in physics

A law in physics is a very useful tool. If applied properly, it enables us to make a very strong argument that what we say is true. If asked 'will a box move?' you can say that you think it will and someone else could say it won't. You both have your opinions and you would then argue as to who is right. However, if you say that Newton's law says it will move, then you have a much stronger argument (assuming you have applied the law correctly).

Newton's first law

A body will remain at rest or moving with constant velocity unless acted upon by an unbalanced force.

To put this the other way round, if the forces on a body are unbalanced, then it will not be at rest or moving with constant velocity. If the velocity is not constant then it is accelerating.

We know from experience that things don't start moving unless we push them, but it's not obvious from observation that things will continue moving with constant velocity unless acted upon by an unbalanced force. Usually, if you give an object a push it moves for a bit and then stops. This is because of friction. It would be a very different world without friction, as everyone would be gliding around with constant velocity, only able to stop themselves when they grabbed hold of something. Friction not only stops things moving but enables them to get going. If you stood in the middle of a friction-free room, you wouldn't be able to move. It is the friction between your feet and the floor that pushes you forward when you try to move your feet backwards.

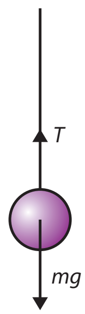


Figure 2.39

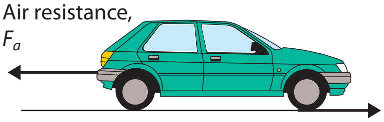


Figure 2.40 Notice that the friction acts forwards, this is because the wheels are trying to turn backwards and friction resists this motion by acting forwards.

Examples

1 **Mass on a string**

If a mass is hanging at rest on the end of a string as in Figure 2.39 then Newton's first law says the forces must be balanced. This means the Force up = Force down.

$T = mg$

2 **Car travelling at constant velocity**

If the car in Figure 2.40 is travelling at constant velocity, then Newton's first law says the forces must be balanced.

Force up = Force down

$N = mg$ (not drawn on diagram)

Force left = Force right

$F = F_a$

3 **The parachutist**

If the free fall parachutist in Figure 2.41 descends at a constant velocity then Newton's first law says that the forces must be balanced.

Force up = Force down

$F_a = mg$

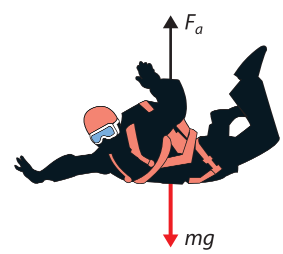


Figure 2.41 A skydiver at terminal velocity.

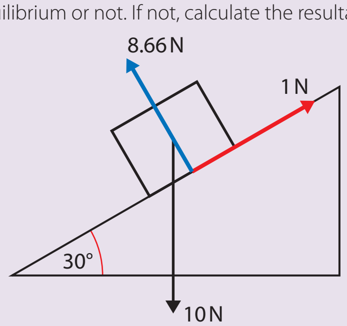
Translational equilibrium

If all the forces on a body are balanced, the body is said to be in translational equilibrium. The bodies in the previous three examples were all therefore in translational equilibrium.

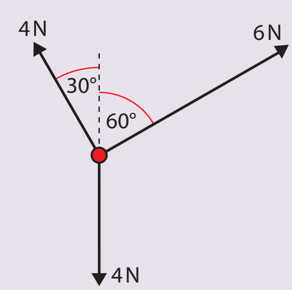
Exercises

22 By resolving the vectors into components, calculate if the following bodies are in translational equilibrium or not. If not, calculate the resultant force.

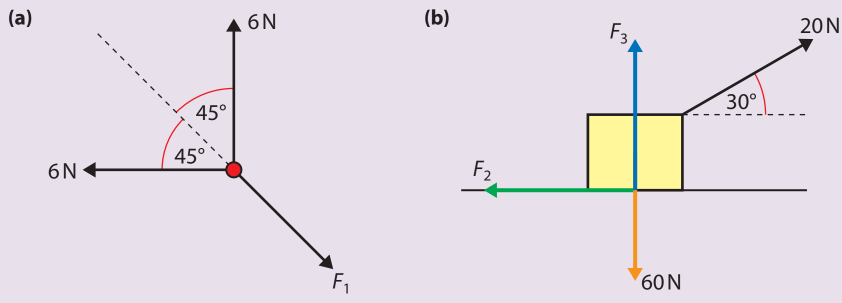
(a)



(b)



23 If the following two examples are in equilibrium, calculate the unknown forces F_1, F_2 and F_3.



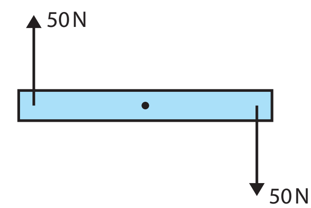


Figure 2.42 The forces are balanced but the body will rotate.

Rotation

When a body is in translational equilibrium it means that its centre will not move. However, it can rotate about the centre, as in the example shown in Figure 2.42.

2.7 The relationship between force and acceleration

Assessment statements

2.2.10 Define linear momentum and impulse.

2.2.8 State Newton's second law of motion.

2.2.9 Solve problems involving Newton's second law.

Newton's first law says that a body will accelerate if an unbalanced force is applied to it. Newton's second law tells us how big the acceleration will be and in which direction. Before we look in detail at Newton's second law we should look at the factors that affect the acceleration of a body when an unbalanced force is applied. Let us consider the example of catching a ball. When we catch the ball we change its velocity, Newton's first law tells us that we must therefore apply an unbalanced force to the ball. The size of that force depends upon two things, the mass and the velocity. A heavy ball is more difficult to stop than a light one travelling at the same speed, and a fast one is harder to stop than a slow one. Rather than having to concern ourselves with two quantities we will introduce a new quantity that incorporates both mass and velocity, *momentum*.

Momentum (p)

Momentum is defined as the product of mass and velocity.

$$p = mv$$

The unit of momentum is kg m s^{-1}.

Momentum is a vector quantity.

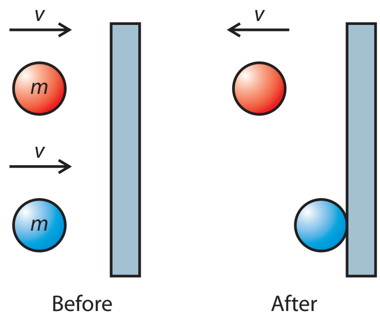


Figure 2.43 The change of momentum of the red ball is greater.

Impulse

When you get hit by a ball the effect it has on you is greater if the ball bounces off you than if you catch it. This is because the change of momentum is greater when the ball bounces, as shown in Figure 2.43.

The unit of impulse is kg m s^{-1}.

Impulse is a vector.

Red ball

Momentum before $= mv$

Momentum after $= -mv$ (remember momentum is a vector)

Change in momentum $= -mv - mv = -2mv$

Blue ball

Momentum before $= mv$

Momentum after $= 0$

Change in momentum $= 0 - mv = -mv$

The impulse is defined as the change of momentum.

Exercises

24 A ball of mass 200 g travelling at 10 m s^{-1} bounces off a wall. If after hitting the wall it travels at 5 m s^{-1}, what is the impulse?

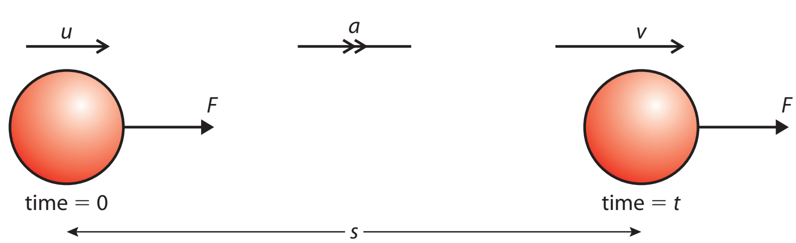
25 Calculate the impulse on a tennis racket that hits a ball of mass 67 g travelling at 10 m s^{-1} so that is comes off the racket at a velocity of 50 m s^{-1}.

Newton's second law

The rate of change of momentum of a body is directly proportional to the unbalanced force acting on that body and takes place in same direction.

Let us once again consider the ball with a constant force acting on it as in Figure 2.44.

Figure 2.44



Unit of momentum

If F = change in momentum /time then momentum = force × time So the unit of momentum is N s. This is the same as kg m s^{-1}.

Firstly Newton's first law tells us that there must be an unbalanced force acting on the ball since it is accelerating.

Newton's second law tells us that the size of the unbalanced force is directly proportional to the rate of change of momentum. We know that the acceleration is constant, which means the rate of change of velocity is constant; this implies that the rate of change of momentum is also constant, so the force, F must be constant too.

If the ball has mass m we can calculate the change of momentum of the ball.

Initial momentum $= mu$

Final momentum $= mv$

Change in momentum $= mv - mu$

The time taken is t so the rate of change of momentum $= \frac{mv - mu}{t}$

This is the same as $\frac{m(v - u)}{t} = ma$

Newton's second law says that the rate of change of momentum is proportional to the force, so $F \propto ma$.

To make things simple the newton is defined so that the constant of proportionality is equal to 1 so:

$$F = ma$$

So when a force is applied to a body in this way, Newton's second law can be simplified to:

The acceleration of a body is proportional to the force applied and inversely proportional to its mass.

Not all examples are so simple. Consider a jet of water hitting a wall as in Figure 2.45. The water hits the wall and loses its momentum, ending up in a puddle on the floor.

Newton's first law tells us that since the velocity of the water is changing, there must be a force on the water,

Newton's second law tells us that the size of the force is equal to the rate of change of momentum. The rate of change of momentum in this case is equal to the amount of water hitting the wall per second multiplied by the change in velocity; this is not the same as ma. For this reason it is best to use the first, more general statement of Newton's second law, since this can always be applied.

However, in this course most of the examples will be of the $F = ma$ type.

The newton

The fact that a 100 g mass has a weight of approximately 1 N is coincidental; 1 N is actually defined as the force that would cause a 1 kg body to accelerate at 1 ms^{-2}.

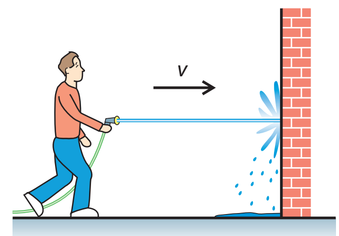


Figure 2.45

Examples

1. Elevator accelerating upwards

An elevator has an upward acceleration of 1 m s^{-2}. If the mass of the elevator is 500 kg, what is the tension in the cables pulling it up?

First draw a free body diagram as in Figure 2.46. Now we can see what forces are acting. Newton's first law tells us that the forces must be unbalanced. Newton's second law tells us that the unbalanced force must be in the direction of the acceleration (upwards). This means that T is bigger than mg.

Newton's second law also tells us that the size of the unbalanced force equals ma so we get the equation

$$T - mg = ma$$

Rearranging gives

$$\begin{aligned} T &= mg + ma \\ &= 500 \times 10 + 500 \times 1 \\ &= 5500\,\text{N} \end{aligned}$$

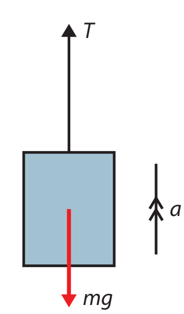


Figure 2.46 An elevator accelerating up. This could either be going up getting faster or going down getting slower.

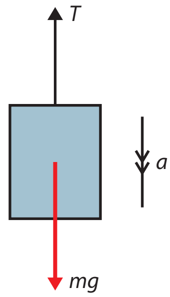


Figure 2.47 The elevator with downward acceleration.

2. Elevator accelerating down

The same elevator as in example 1 now has a downward acceleration of 1 m s^{-2} as in Figure 2.47.

This time Newton's laws tell us that the weight is bigger than the tension so $mg - T = ma$

Rearranging gives

$$\begin{aligned} T &= mg - ma \\ &= 500 \times 10 - 500 \times 1 \\ &= 4500\,\text{N} \end{aligned}$$

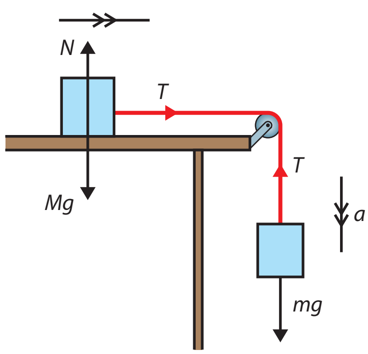


Figure 2.48

3. Joined masses

Two masses are joined by a rope. One of the masses sits on a frictionless table, the other hangs off the edge as in Figure 2.48.

M is being dragged to the edge of the table by m.

Both are connected to the same rope so T is the same for both masses, this also means that the acceleration a is the same.

We do not need to consider N and Mg for the mass on the table because these forces are balanced. However the horizontally unbalanced force is T.

Applying Newton's laws to the mass on the table gives

$$T = Ma$$

The hanging mass is accelerating down so mg is bigger than T. Newton's second law implies that $mg - T = ma$

Substituting for T gives $mg - Ma = ma$ so $a = \dfrac{mg}{M + m}$

4. The free fall parachutist

After falling freely for some time, a free fall parachutist whose weight is 60 kg opens her parachute. Suddenly the force due to air resistance increases to 1200 N. What happens?

Looking at the free body diagram in Figure 2.49 we can see that the forces are unbalanced and that according to Newton's second law the acceleration, a, will be upwards.

The size of the acceleration is given by

$$ma = 1200 - 600 = 60 \times a$$

so $$a = 10\,\text{m s}^{-2}$$

The acceleration is in the opposite direction to the motion. This will cause the parachutist to slow down. As she slows down, the air resistance gets less until the forces are balanced. She will then continue down with a constant velocity.

Figure 2.49 The parachutist just after opening the parachute.

Exercise

26 The helium in a balloon causes an upthrust of 0.1 N. If the mass of the balloon and helium is 6 g, calculate the acceleration of the balloon.

27 A rope is used to pull a felled tree (mass 50 kg) along the ground. A tension of 1000 N causes the tree to move from rest to a velocity of 0.1 m s^{-1} in 2 s. Calculate the force due to friction acting on the tree.

28 Two masses are arranged on a frictionless table as shown in Figure 2.50. Calculate:

(a) the acceleration of the masses

(b) the tension in the string.

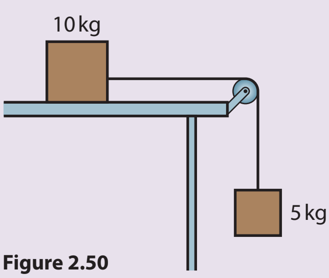


Figure 2.50

29 A helicopter is lifting a load of mass 1000 kg with a rope. The rope is strong enough to hold a force of 12 kN. What is the maximum upward acceleration of the helicopter?

30 A person of mass 65 kg is standing in an elevator that is accelerating upwards at 0.5 m s^{-2}. What is the normal force between the floor and the person?

31 A plastic ball is held under the water by a child in a swimming pool. The volume of the ball is 4000 cm^3.

(a) If the density of water is 1000 kg m^{-3}, calculate the upthrust on the ball (remember upthrust = **weight** of fluid displaced).

(b) If the mass of the ball is 250 g, calculate the theoretical acceleration of the ball when it is released. Why won't the ball accelerate this quickly in a real situation?

2.8 Newton's third law

Assessment statements

2.2.14 State Newton's third law of motion.
2.2.15 Discuss examples of Newton's third law.
2.2.12 State the law of conservation of linear momentum.
2.2.13 Solve problems involving momentum and impulse.
2.2.11 Determine the impulse due to a time–varying force by interpreting a force–time graph.

When dealing with Newton's first and second laws, we are careful to consider only the body that is *experiencing* the forces, not the body that is *exerting* the forces. Newton's third law relates these forces.

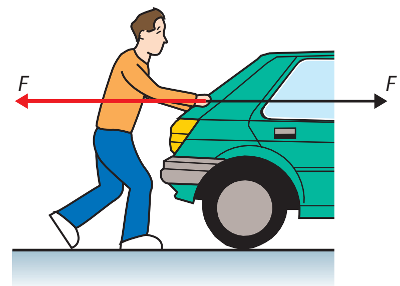
If body A exerts a force on body B then body B will exert an equal and opposite force on body A.

Incorrect statements

It is very important to realise that Newton's third law is about two bodies. Avoid statements of this law that do not mention anything about there being two bodies.

So if someone is pushing a car with a force F as shown in Figure 2.51 the car will push back on the person with a force $-F$. In this case both of these forces are the normal force.

Figure 2.51 The man pushes the car and the car pushes the man.



You might think that since these forces are equal and opposite, they will be balanced, and in that case how does the person get the car moving? This is wrong; the forces act on different bodies so can't balance each other.

Examples

1. A falling body

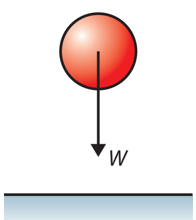


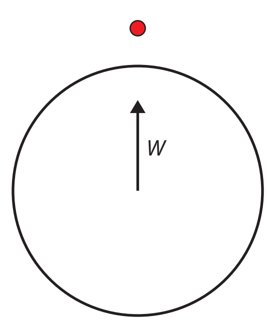
Figure 2.52 A falling body pulled down by gravity.

A body falls freely towards the ground as in Figure 2.52. If we ignore air resistance, there is only one force acting on the body – the force due to the gravitational attraction of the Earth, that we call weight.

Applying Newton's third law:

If the Earth pulls the body down, then the body must pull the Earth up with an equal and opposite force. We have seen that the gravitational force always acts on the centre of the body, so Newton's third law implies that there must be a force equal to W acting upwards on the centre of the Earth as in Figure 2.53.

Figure 2.53 The Earth pulled up by gravity.



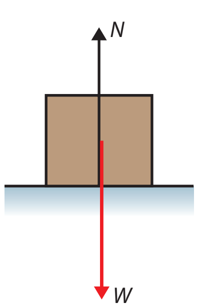


Figure 2.54 Forces acting on a box resting on the floor.

2. A box rests on the floor

A box sits on the floor as shown in Figure 2.54. Let us apply Newton's third law to this situation.

There are two forces acting on the box.

Normal force: The floor is pushing up on the box with a force N. According to Newton's third law the box must therefore push down on the floor with a force of magnitude N.

Weight: The Earth is pulling the box down with a force W. According to Newton's third law, the box must be pulling the Earth up with a force of magnitude W as shown in Figure 2.55.

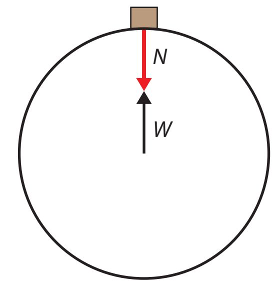


Figure 2.55 Forces acting on the Earth according to Newton's third law.

3. Recoil of a gun

When a gun is fired the velocity of the bullet changes. Newton's first law implies that there must be an unbalanced force on the bullet; this force must come from the gun. Newton's third law says that if the gun exerts a force on the bullet the bullet must exert an equal and opposite force on the gun. This is the force that makes the gun recoil or 'kick back'.

4. The water cannon

When water is sprayed at a wall from a hosepipe it hits the wall and stops. Newton's first law says that if the velocity of the water changes, there must be an unbalanced force on the water. This force comes from the wall. Newton's third law says that if the wall exerts a force on the water then the water will exert a force on the wall. This is the force that makes a water cannon so effective at dispersing demonstrators.

A boat tests its water cannons.

Exercise

32 Use Newton's first and third laws to explain the following:

(a) When burning gas is forced downwards out of a rocket motor, the rocket accelerates up.

(b) When the water cannons on the boat in the photo are operating, the boat accelerates forwards.

(c) When you step forwards off a skateboard, the skateboard accelerates backwards.

(d) A table tennis ball is immersed in a fluid and held down by a string as shown in Figure 2.56. The container is placed on a balance. What will happen to the reading of the balance if the string breaks?

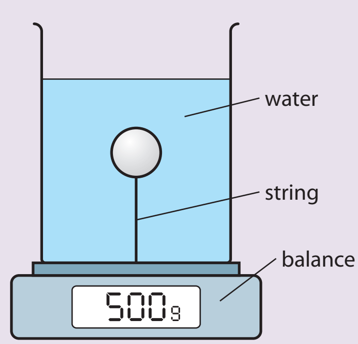
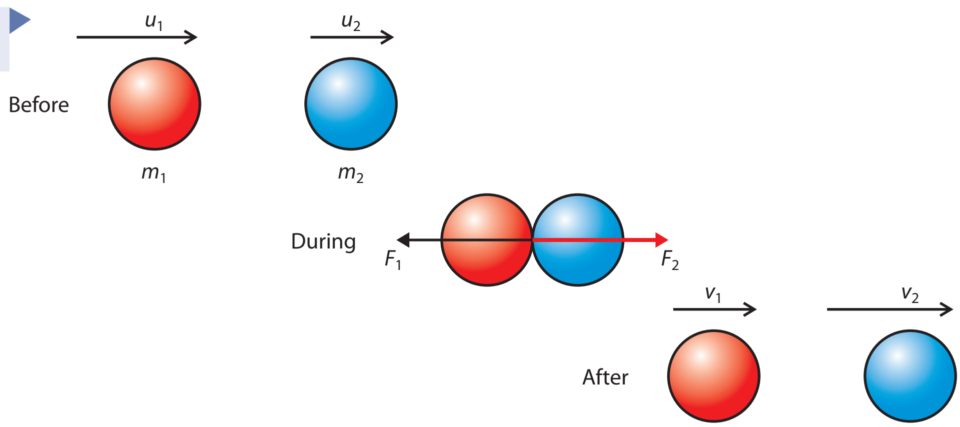


Figure 2.56

Collisions

In this section we have been dealing with the interaction between two bodies (gun-bullet, skater-skateboard, hose-water). To develop our understanding of the interaction between bodies, let us consider a simple collision between two balls as illustrated in Figure 2.57.

Figure 2.57 Collision between two balls.



Let us apply Newton's three laws to this problem.

Newton's first law

In the collision the red ball slows down and the blue ball speeds up. Newton's first law tells us that that this means there is a force acting to the left on the red ball (F_1) and to the right on the blue ball (F_2).

Newton's second law

This law tells us that the force will be equal to the rate of change of momentum of the balls so if the balls are touching each other for a time Δt

$$F_1 = \frac{m_1 v_1 - m_1 u_1}{\Delta t}$$

$$F_2 = \frac{m_2 v_2 - m_2 u_2}{\Delta t}$$

Isolated system
An isolated system is one in which no external forces are acting. When a ball hits a wall the momentum of the ball is not conserved because the ball and wall is not an isolated system, since the wall is attached to the ground. If the ball and wall were floating in space then momentum would be conserved.

Newton's third law

According to the third law, if the red ball exerts a force on the blue ball, then the blue ball will exert an equal and opposite force on the red ball.

$$F_1 = -F_2$$

$$\frac{m_1 v_1 - m_1 u_1}{\Delta t} = \frac{-m_2 v_2 - m_2 u_2}{\Delta t}$$

Rearranging gives $m_1 u_1 + m_2 u_2 = m_1 v_1 + m_2 v_2$

In other words the momentum at the start equals the momentum at the end. We find that this applies not only to this example but to all interactions.

The law of the conservation of momentum

For a system of isolated bodies the total momentum is always the same.

This is not a new law since it is really just a combination of Newton's laws. However it provides a useful short cut when solving problems.

Examples

In these examples we will have to pretend everything is in space isolated from the rest of the universe, otherwise they are not isolated and the law of conservation of momentum won't apply.

1. A collision where the bodies join together

If two balls of modelling clay collide with each other they stick together as shown in Figure 2.58. We want to find the velocity, v, of the combined lump after the collision.

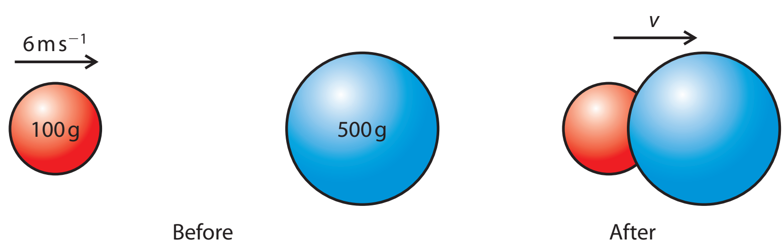


Figure 2.58 Two bodies stick together after colliding.

Simplified models
Pieces of clay floating in space are not exactly everyday examples, but most everyday examples (like balls on a pool table) are not isolated systems, so we can't solve them in this simple way.

If bodies are isolated then momentum is conserved so:

$$\text{momentum before} = \text{momentum after}$$
$$0.1 \times 6 + 0.5 \times 0.0 = 0.6 \times v$$
$$v = \frac{0.6}{0.6} = 1\ \text{m s}^{-1}$$

2. An explosion

A ball of clay floating around in space suddenly explodes into a big piece and a small piece, as shown in Figure 2.59. If the big bit has a velocity of $5\ \text{m s}^{-1}$, what is the velocity of the small bit?

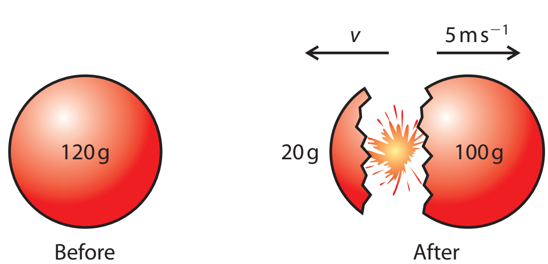


Figure 2.59 A piece of modelling clay suddenly explodes.

Since this is an isolated system, momentum is conserved so:

$$\text{momentum before} = \text{momentum after}$$
$$0 \times 0.12 = 0.02 \times (-v) + 0.1 \times 5$$
$$0.02 \times v = 0.5$$
$$v = 25\ \text{m s}^{-1}$$

Exercise

33 Draw diagrams to represent the following collisions then use the law of conservation of momentum to find the unknown velocity. Assume all collisions are head-on, in other words they take place in one dimension.

(a) Two identical isolated balls collide with each other. Before the collision, one ball was travelling at $10\ \mathrm{m\ s^{-1}}$ and the other was at rest. After the collision the first ball continues in the same direction with a velocity of $1\ \mathrm{m\ s^{-1}}$. Find the velocity of the other ball.

(b) Two identical balls are travelling towards each other; each is travelling at a speed of $5\ \mathrm{m\ s^{-1}}$. After they hit, one ball bounces off with a speed of $1\ \mathrm{m\ s^{-1}}$. What is the speed of the other?

(c) A spaceman of mass 100 kg is stranded 2 m from his spaceship as shown in Figure 2.60. He happens to be holding a hammer of mass 2 kg what must he do?

If he only has enough air to survive for 2 minutes, how fast must he throw the hammer if he is to get back in time? Is it possible?

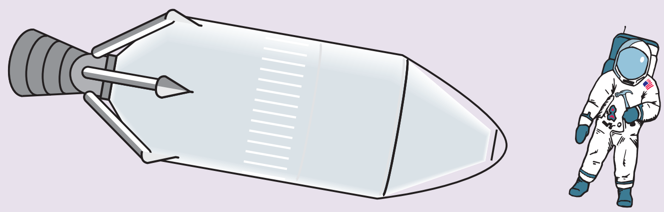


Figure 2.60 If you are ever in this position this course could save your life.

What happens during the collision?

In the previous examples, we were more interested in the difference between the conditions before and after the collision than during it. However, collisions can be very different. For example, consider a ball of mass 200 g colliding with a hard floor and a trampoline as shown in Figure 2.61. Before the collisions each ball travels downwards at $10\ \mathrm{m\ s^{-1}}$ and each bounces up with velocity $10\ \mathrm{m\ s^{-1}}$. So the change in momentum (impulse) is the same for each:

$$0.2 \times (-20) - 0.2 \times 20 = -8\ \mathrm{N\ s}$$

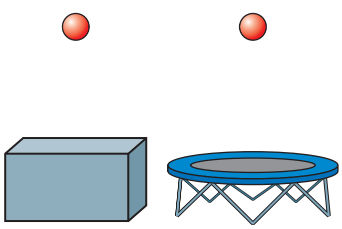


Figure 2.61 It's less painful to land on a trampoline than a concrete block.

Each has the same change of momentum but each collision was very different – the collision with the trampoline took place over a longer time. If you replace the ball with yourself, you would certainly be able to feel the difference. We can represent these two collisions by plotting a graph of force against time as shown in Figure 2.62. We see the force exerted by the concrete is much greater.

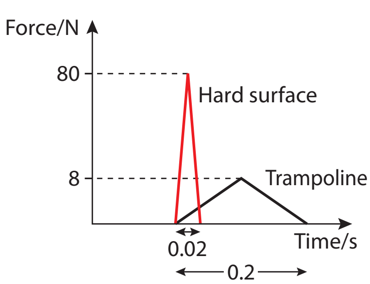


Figure 2.62

Area under the graph

We notice that the area under the graph for each interaction is the same.

Area $= \frac{1}{2}$base $\times$ height $= \frac{1}{2} \times 0.02 \times 8 = \frac{1}{2} \times 0.2 \times 0.8 = 8\ \mathrm{N\ s}$

This is equal to the impulse.

Exercise

34 (a) Calculate the impulse of the body for the motion represented in Figure 2.63.

(b) If the mass of the object is 20 g, what is the change of velocity?

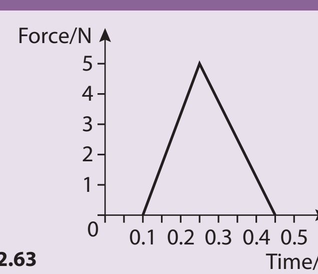


Figure 2.63

2.9 Work, energy and power

Assessment statements

2.3.1 Outline what is meant by work.
2.3.2 Determine the work done by a non-constant force by interpreting a force–displacement graph.
2.3.3 Solve problems involving the work done by a force.
2.3.4 Outline what is meant by kinetic energy.
2.3.5 Outline what is meant by change in gravitational potential energy.
2.3.6 State the principle of conservation of energy.
2.3.7 List different forms of energy and describe examples of the transformation of energy from one form to another.
2.3.8 Distinguish between elastic and inelastic collisions.
2.3.9 Define power.
2.3.10 Define and apply the concept of efficiency.
2.3.11 Solve problems involving momentum, work, energy and power.

We have so far dealt with the motion of a small red ball and understand what causes it to accelerate. We have also investigated the interaction between a red ball and a blue one and have seen that the red one can cause the blue one to move when they collide. But what enables the red one to push the blue one? To answer this question we need to define some more quantities.

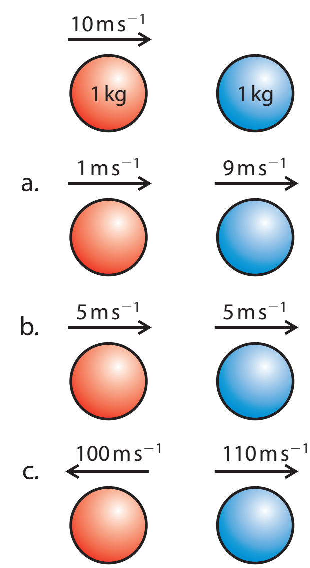


Figure 2.64 The red ball hits the blue ball but what happens?

Work

In the introduction to this book it was stated that by developing models, our aim is to understand the physical world so that we can make predictions. At this point you should understand certain concepts related to the collision between two balls, but we still can't predict the outcome. To illustrate this point let us again consider the red and blue balls. Figure 2.64 shows three possible outcomes of the collision.

If we apply the law of conservation of momentum, we realise that all three outcomes are possible. The original momentum is 10 newtonmetres (10 Nm) and the final momentum is 10 Nm in all three cases. But which one actually happens? This we cannot say (yet). All we know is that from experience the last option is not possible – but why?

When body A hits body B, body A exerts a force on body B. This force causes B to have an increase in velocity. The amount that the velocity increases depends upon how big the force is and over what distance the collision takes place. To make this simpler, consider a constant force acting on a body as in Figure 2.65.

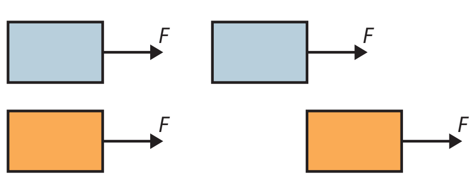


Figure 2.65 The force acts on the orange block for a greater distance.

Both blocks start at rest and are pulled by the same force, but the second block will gain more velocity because the force acts over a longer distance. To quantify this difference, we say that in the second case the force has done more work. Work is done when the point of application of a force moves in the direction of the force.

Work is defined in the following way:

Work done = force × distance moved in the direction of the force.

The unit of work is the newtonmetre (Nm) which is the same as the joule (J).

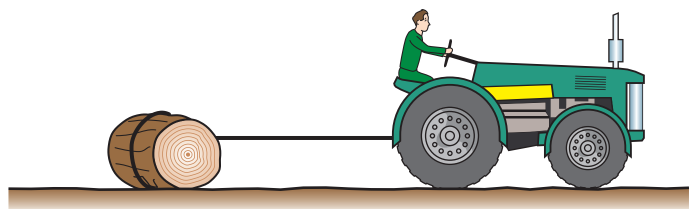
Work is a scalar quantity.

Sign of work

Since work is a scalar the sign has nothing to do with direction. Saying that you have done negative work on A is the same as saying A has done work on you.

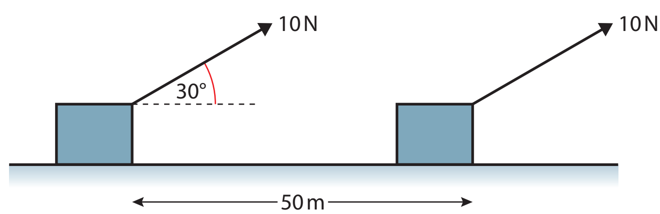
Worked examples

1 A tractor pulls a felled tree along the ground for a distance of 200 m. If the tractor exerts a force of 5000 N, how much work will be done?



2 A force of 10 N is applied to a block, pulling it 50 m along the ground as shown in Figure 2.66. How much work is done by the force?

Figure 2.66



3 When a car brakes it slows down due to the friction force between the tyres and the road. This force opposes the motion as shown in Figure 2.67. If the friction force is a constant 500 N and the car comes to rest in 25 m, how much work is done by the friction force?

Figure 2.67 Work done against friction.

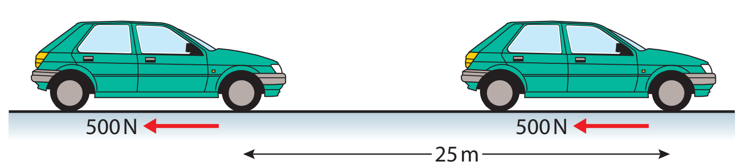
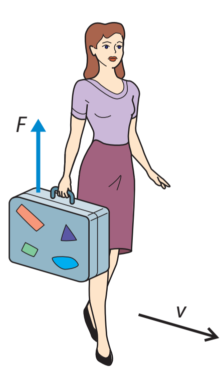


Figure 2.68



4 The woman in Figure 2.68 walks along with a constant velocity holding a suitcase. How much work is done by the force holding the case?

Solutions

1 Work done = force × distance moved in direction of force

Work done = 5000 × 200 = 1 MJ

2 In this example the force is not in the same direction as the movement. However, the horizontal component of the force is.

Work done = 10 × cos 30° × 50 = 433 N

3 This time the force is in the opposite direction to the motion.

Work done = −500 × 25 = −12 500 J

The negative sign tells us that the friction isn't doing the work but the work is being done against the friction.

4 In this example the force is acting perpendicular to the direction of motion, so there is no movement in the direction of the force.

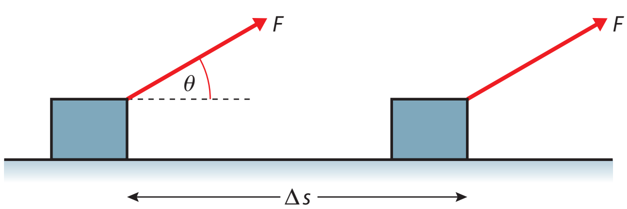
Work done = zero

General formula

In general

$$\text{Work} = F\cos\theta \times \Delta s$$

where θ is the angle between the displacement, Δs, and force, F.



All the previous examples can be solved using this formula.

If $\theta < 90°$, $\cos\theta$ is positive so the work is positive.

$\theta = 90°$, $\cos\theta = 0$ so the work is zero.

$\theta > 90°$, $\cos\theta$ is negative so the work is negative.

Working or not?
It may seem strange that when you carry a heavy bag you are not doing any work – that's not what it feels like. In reality, lots of work is being done, since to hold the bag you use your muscles. Muscles are made of microscopic fibres, which are continuously contracting and relaxing, so are doing work.

Exercises

35 Figure 2.69 shows a boy taking a dog for a walk.
(a) Calculate the work done by the force shown when the dog moves 10 m forward.
(b) Who is doing the work?

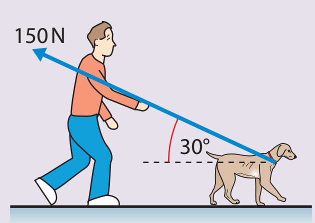


Figure 2.69

36 A bird weighing 200 g sits on a tree branch.
How much work does the bird do on the tree?

37 As a box slides along the floor it is slowed down by a constant force due to friction. If this force is 150 N and the box slides for 2 m, how much work is done against the frictional force?

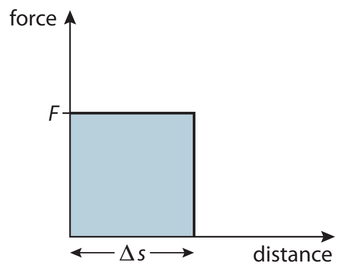


Figure 2.70 Force *vs* distance for a constant force.

Work done by a varying force

In the examples so far, the forces have been constant. If the force isn't constant, we can't simply say work = force × distance unless we use the 'average force'. In these cases, we can use a graphical method.

Let us first consider a constant force, F, acting over a distance Δs. The graph of force against distance for this motion is shown in Figure 2.70.

From the definition of work, we know that in this case work $= F\Delta s$

This is the same as the area under the graph. From this we deduce that:

$$\text{work done} = \text{area under the force–distance graph}$$

Example

Stretching a spring

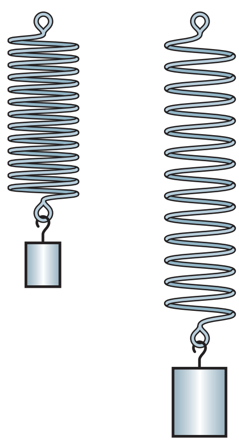


Figure 2.71 Stretching a spring.

Stretching a spring is a common example of a varying force. When you stretch a spring it gets more and more difficult the longer it gets. Within certain limits the force needed to stretch the spring is directly proportional to the extension of the spring. This was first recognised by Robert Hooke in 1676, so is named 'Hooke's Law'. Figure 2.71 shows what happens if we add different weights to a spring; the more weight we add the longer it gets. If we draw a graph of force against distance as we stretch a spring, it will look like the graph in Figure 2.72. The gradient of this line, $\frac{F}{\Delta s}$ is called the spring constant, k.

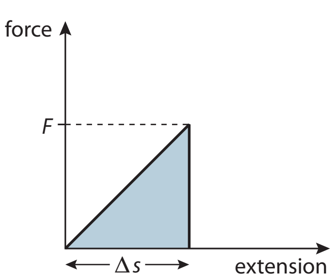


Figure 2.72 Force *vs* extension for a spring.

The work done as the spring is stretched is found by calculating the area under the graph.

$$\text{area} = \tfrac{1}{2}\text{base} \times \text{height} = \tfrac{1}{2}F\Delta s$$

So $\quad \text{work done} = \tfrac{1}{2}F\Delta s$

But if $\quad \frac{F}{\Delta s} = k$ then $F = k\Delta s$

Substituting for F gives

$$\text{work done} = \tfrac{1}{2}k\Delta s^2$$

Exercises

38 A spring of spring constant 2 N cm^{-1} and length 6 cm is stretched to a new length of 8 cm.

(a) How far has the spring been stretched?

(b) What force will be needed to hold the spring at this length?

(c) Sketch a graph of force against extension for this spring.

(d) Calculate the work done in stretching the spring.

(e) The spring is now stretched a further 2 cm. Draw a line on your graph to represent this and calculate how much additional work has been done.

39 Calculate the work done by the force represented by Figure 2.73.

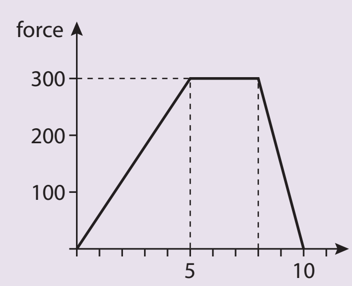


Figure 2.73

Energy

We have seen that it is sometimes possible for body A to do work on body B but what does A have that enables it to do work on B? To answer this question we must define a new quantity, energy.

Energy is the quantity that enables body A to do work on body B.

If body A collides with body B as shown in Figure 2.74, body A has done work on body B. This means that body B can now do work on body C. Energy has been transferred from A to B.

When body A does work on body B, energy is transferred from body A to body B.

The unit of energy is the joule (J).

Energy is a scalar quantity.

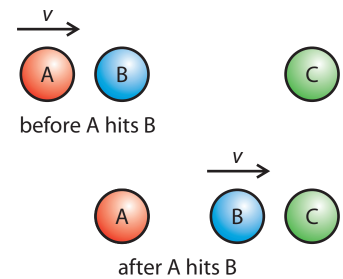


Figure 2.74 The red ball gives energy to the blue ball.

Different types of energy

If a body can do work then it has energy. There are two ways that a simple body such as a red ball can do work. In the example above, body A could do work because it was moving – this is called *kinetic energy*. Figure 2.75 shows an example where A can do work even though it isn't moving. In this example, body A is able to do work on body B because of its position above the Earth. If the hand is removed, body A will be pulled downwards by the force of gravity, and the string attached to it will then drag B along the table. If a body is able to do work because of its position, we say it has *potential energy*.

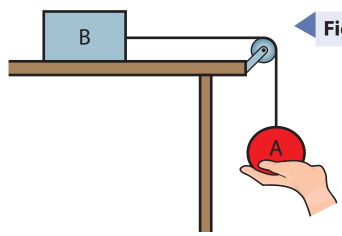


Figure 2.75

Kinetic energy (KE)

This is the energy a body has due to its movement. To give a body KE, work must be done on the body. The amount of work done will be equal to the increase in KE. If a constant force acts on a red ball of mass m as shown in Figure 2.76, then the work done is Fs.

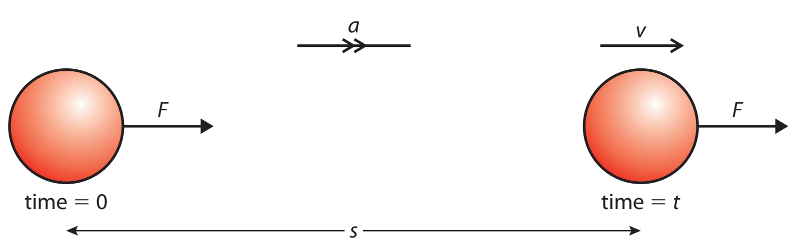


Figure 2.76

From Newton's second law we know that $F = ma$ which we can substitute in work $= Fs$ to give work $= mas$

We also know that since acceleration is constant we can use the *suvat* equation $v^2 = u^2 + 2as$ which since $u = 0$ simplifies to $v^2 = 2as$.

Rearranging this gives $as = \frac{v^2}{2}$ so work $= \frac{1}{2}mv^2$

This work has increased the KE of the body so we can deduce that:

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

Use of words

If we say a body has potential energy it sounds as though it has the potential to do work. This is true, but a body that is moving has the potential to do work too. This can lead to misunderstanding. It would have been better to call it *positional energy*.

Other types of PE
In this section we only deal with examples of PE due to a body's position close to the Earth. However there are other positions that will enable a body to do work (for example, in an electric field). These will be introduced after the concept of fields has been introduced in Chapter 6.

Gravitational potential energy (PE)

This is the energy a body has due to its position above the Earth.

For a body to have PE it must have at some time been lifted to that position. The amount of work done in lifting it equals the PE. Taking the example shown in Figure 2.75, the work done in lifting the mass, m, to a height h is mgh (this assumes that the body is moving at a constant velocity so the lifting force and weight are balanced).

If work is done on the body then energy is transferred so:

gain in PE $= mgh$

The law of conservation of energy

We could not have derived the equations for KE or PE without assuming that the work done is the same as the gain in energy. The law of conservation of energy is a formal statement of this fact.

Energy can neither be created nor destroyed – it can only be changed from one form to another.

This law is one of the most important laws that we use in physics. If it were not true you could suddenly find yourself at the top of the stairs without having done any work in climbing them, or a car suddenly has a speed of 200 km h^{-1} without anyone touching the accelerator pedal. These things just don't happen, so the laws we use to describe the physical world should reflect that.

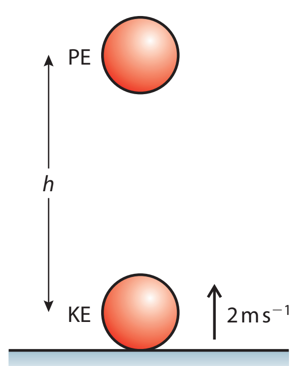


Figure 2.77 Work is done lifting the ball so it gains PE.

Worked examples

1 A ball of mass 200 g is thrown vertically upwards with a velocity of 2 m s^{-1} as shown in Figure 2.77. Use the law of conservation of energy to calculate its maximum height.

2 A block slides down the frictionless ramp shown in Figure 2.78. Use the law of conservation of energy to find its speed when it gets to the bottom.

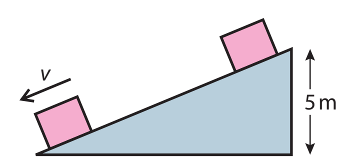


Figure 2.78 As the block slides down the slope it gains KE.

Solutions

1 At the start of its motion the body has KE. This enables the body to do work against gravity as the ball travels upwards. When the ball reaches the top, all the KE has been converted into PE. So applying the law of conservation of energy:

$$\text{loss of KE} = \text{gain in PE}$$

$$\tfrac{1}{2}mv^2 = mgh$$

so $$h = \frac{v^2}{2g} = \frac{2^2}{2 \times 10} = 0.2\ \text{m}$$

This is exactly the same answer you would get by calculating the acceleration from $F = ma$ and using the *suvat* equations.

2 This time the body loses PE and gains KE so applying the law of conservation of energy:

$$\text{loss of PE} = \text{gain of KE}$$

$$mgh = \tfrac{1}{2}mv^2$$

So $$v = \sqrt{2gh} = \sqrt{(2 \times 10 \times 5)} = 10\,\text{m s}^{-1}$$

Again, this is a much simpler way of getting the answer than using components of the forces.

Exercises

Use the law of conservation of energy to solve the following:

40 A stone of mass 500 g is thrown off the top of a cliff with a speed of $5\,\text{m s}^{-1}$. If the cliff is 50 m high, what is its speed just before it hits the ground?

41 A ball of mass 250 g is dropped 5 m onto a spring as shown in Figure 2.79.
- **(a)** How much KE will the ball have when it hits the spring?
- **(b)** How much work will be done as the spring is compressed?
- **(c)** If the spring constant is $250\,\text{kN m}^{-1}$, calculate how far the spring will be compressed.

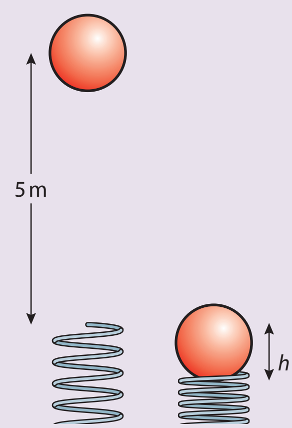


Figure 2.79 In this example the spring is compressed, not stretched, but Hooke's law still applies.

42 A ball of mass 100 g is hit vertically upwards with a bat. The bat exerts a constant force of 15 N on the ball and is in contact with it for a distance of 5 cm.
- **(a)** How much work does the bat do on the ball?
- **(b)** How high will the ball go?

43 A child pushes a toy car of mass 200 g up a slope. The car has a speed of $2\,\text{m s}^{-1}$ at the bottom of the slope.
- **(a)** How high up the slope will the car go?
- **(b)** If the speed of the car were doubled how high would it go now?

Forms of energy

When we are describing the motion of simple red balls there are only two forms of energy, KE and PE. However, when we start to look at more complicated systems, we discover that we can do work using a variety of different machines, such as petrol engine, electric engine etc. To do work, these machines must be given energy and this can come in many forms, for example:

- Petrol
- Gas
- Electricity
- Solar
- Nuclear

As you learn more about the nature of matter in Chapter 3, you will discover that all of these (except solar) are related to either KE or PE of particles.

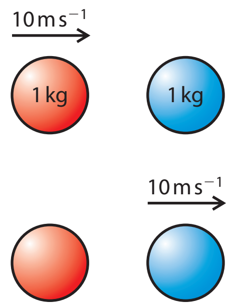


Figure 2.80

Energy and collisions

One of the reasons that we brought up the concept of energy was related to the collision between two balls as shown in Figure 2.80. We now know that if no energy is lost when the balls collide, then the KE before the collision = KE after. This enables us to calculate the velocity afterwards and the only solution in this example is quite a simple one. The red ball gives all its KE to the blue one, so the red one stops and the blue one continues, with velocity = $10\,\mathrm{m\,s^{-1}}$. If the balls become squashed, then some work needs to be done to squash them. In this case not all the KE is transferred, and we can only calculate the outcome if we know how much energy is used in squashing the balls.

Elastic collisions

An elastic collision is a collision in which both momentum and KE are conserved.

Example

Two balls of equal mass collide with each other elastically with the velocities shown in Figure 2.81. What is the velocity of the balls after the collision?

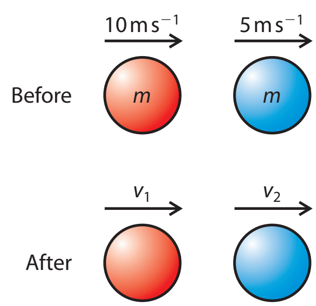


Figure 2.81

● **Examiner's hint:** If the masses are the same in an elastic collision the velocities of the two bodies swap

Conservation of momentum:

Momentum before = momentum after

$m \times 10 + m \times 5 = mv_1 + mv_2$

$10 + 5 = v_1 + v_2$

Conservation of KE:

$\frac{1}{2}m \times 10^2 + \frac{1}{2}m \times 5^2$

$= \frac{1}{2}mv_1^2 + \frac{1}{2}mv_2^2$

$10^2 + 5^2 = v_1^2 + v_2^2$

There are only two possible solutions to these two equations; either the velocities don't change, which means there isn't a collision, or the velocities swap, so $v_1 = 5\,\mathrm{m\,s^{-1}}$ and $v_2 = 10\,\mathrm{m\,s^{-1}}$.

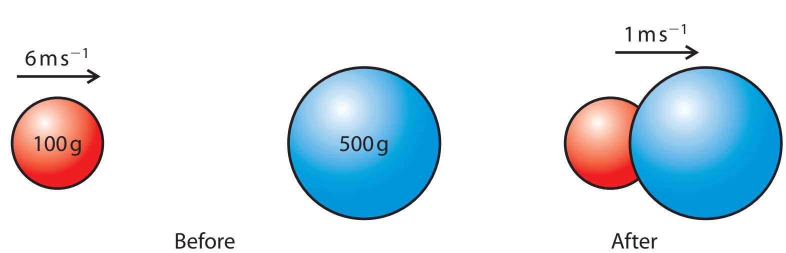
Inelastic collisions

There are many outcomes of an inelastic collision but here we will only consider the case when the two bodies stick together. We call this *totally inelastic collision*.

Example

When considering the conservation of momentum in collisions, we used the example shown in Figure 2.82. How much work was done to squash the balls in this example?

Figure 2.82



According to the law of conservation of energy, the work done squashing the balls is equal to the loss in KE.

KE loss = KE before − KE after $= \frac{1}{2} \times 0.1 \times 6^2 - \frac{1}{2} \times 0.6 \times 1^2$

KE loss $= 1.8 - 0.3 = 1.5\,\mathrm{J}$

So work done $= 1.5\,\mathrm{J}$

Explosions

Explosions can never be elastic since, without doing work and therefore increasing the KE, the parts that fly off after the explosion would not have any KE and would therefore not be moving. The energy to initiate an explosion often comes from the chemical energy contained in the explosive.

Example

Again consider a previous example where a ball exploded (shown again in Figure 2.83). How much energy was supplied to the balls by the explosive?

Sharing of energy
The result of this example is very important; we will use it when dealing with nuclear decay later on. So remember, when a body explodes into two unequal bits, the small bit gets most energy.

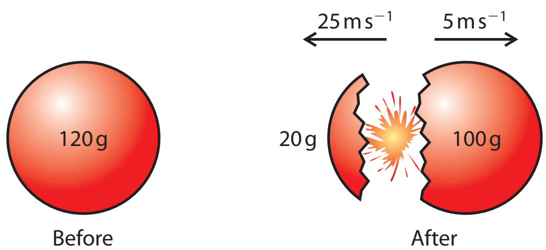


Figure 2.83

According to the law of conservation of energy, the energy from the explosive equals the gain in KE of the balls.

Notice how the small bit gets most of the energy. We will come back to this when studying nuclear decay.

KE gain = KE after − KE before

$$\text{KE gain} = (\tfrac{1}{2} \times 0.02 \times 25^2 + \tfrac{1}{2} \times 0.1 \times 5^2) - 0 = 6.25 + 1.25 = 7.5\ \text{J}$$

Exercise

44 Two balls are held together by a spring as shown in Figure 2.84. The spring has a spring constant of 10 N cm^{-1} and has been compressed a distance 5 cm.

(a) How much work was done to compress the spring?

(b) How much KE will each gain?

(c) If each ball has a mass of 10 g calculate the velocity of each ball.

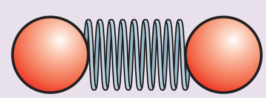


Figure 2.84

45 Two pieces of modelling clay as shown in Figure 2.85 collide and stick together.

(a) Calculate the velocity of the lump after the collision.

(b) How much KE is lost during the collision?

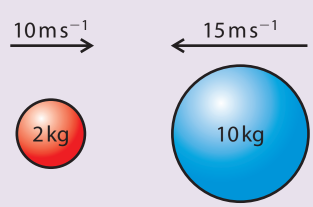


Figure 2.85

46 A red ball travelling at 10 m s^{-1} to the right collides with a blue ball with the same mass travelling at 15 m s^{-1} to the left. If the collision is elastic, what are the velocities of the balls after the collision?

Power

We know that to do work requires energy, but work can be done quickly or it can be done slowly. This does not alter the energy exchanged but the situations are certainly different. For example we know that to lift one thousand, 1 kg bags of sugar from the floor to the table is not an impossible task – we can simply lift them one by one. It will take a long time but we would manage it in the end. However, if we were asked to do the same task in 5 seconds, we would either have

Power and velocity

If power $= \frac{\text{work done}}{\text{time}}$ then we can also write $P = \frac{F\Delta s}{t}$ so $P = F\left(\frac{\Delta s}{t}\right)$ which is the same as $P = Fv$ where v is the velocity.

to lift all 1000 kg at the same time or move each bag in 0.005 s, both of which are impossible. Power is the quantity that distinguishes between these two tasks.

Power is defined as:

Power = work done per unit time.

Unit of power is the J/s which is the same as the watt (W).

Power is a scalar quantity.

Examples

1. The powerful car

We often use the term power to describe cars. A powerful car is one that can accelerate from 0 to 100 km h^{-1} in a very short time. When a car accelerates, energy is being converted from the chemical energy in the fuel to KE. To have a big acceleration the car must gain KE in a short time, hence be powerful.

Horse power

Horse power is often used as the unit for power when talking about cars and boats.

746 W = 1 hp

So in Example 1, the power of the car is 105 hp.

2. Power lifter

A power lifter is someone who can lift heavy weights, so shouldn't we say they are strong people rather than powerful? A power lifter certainly is a strong person (if they are good at it) but they are also powerful. This is because they can lift a big weight in a short time.

Worked example

A car of mass 1000 kg accelerates from rest to 100 km h^{-1} in 5 seconds. What is the average power of the car?

Solution

100 km h^{-1} = 28 m s^{-1}.

The gain in KE of the car $= \frac{1}{2}mv^2 = \frac{1}{2} \times 1000 \times 28^2 = 392$ kJ

If the car does this in 5 s then

$$\text{power} = \frac{\text{work done}}{\text{time}} = \frac{392}{5} = 78.4 \text{ kW}$$

Exercises

47 A weightlifter lifts 200 kg 2 m above the ground in 5 s. Calculate the power of the weightlifter in watts.

48 In 25 s a trolley of mass 50 kg runs down a hill. If the difference in height between the top and the bottom of the hill is 50 m, how much power will have been dissipated?

49 A car moves along a road at a constant velocity of 20 m s^{-1}. If the resistance force acting against the car is 1000 N, what is the power developed by the engine?

Efficiency

Efficiency and power

Since the work is going in and out at the same rate, efficiency can also be expressed as

$$\frac{\text{power out}}{\text{power in}}$$

When a box is pulled along the floor, the person pulling has to do work. This work is converted into the KE of the box and some of it is done against friction. Since the result they are trying to achieve is to move the box, the energy transferred to KE could be termed 'useful energy' and the rest is 'wasted'. (The waste energy in this example turns into heat, but we will deal with that in the next chapter). The efficiency tells us how good a system is at turning the input energy into useful work.

Efficiency is defined as:

$$\frac{\text{useful work out}}{\text{energy put in}}$$

Efficiency has no units.

Efficiency is a scalar quantity.

It is worth noting here that due to the law of conservation of energy, the efficiency can never be greater than 1.

Worked example

A box of mass 10 kg is pulled along the floor for 2 m by a horizontal force of 50 N. If the frictional force is 20 N, what is the efficiency of the system?

Solution

The work done by the pulling force = force × distance = $50 \times 2 = 100$ J

The unbalanced force on the box = $50 - 20 = 30$ N

So the work done on the box = $30 \times 2 = 60$ J

This work is exchanged to the box and increases its KE.

Work done against friction = $20 \times 2 = 40$ J

So work in = 100 J and total work out = $60 + 40 = 100$ J

Efficiency = $\frac{\text{useful work out}}{\text{work in}} = \frac{60}{100} = 0.6$

Expressed as a percentage, this is 60%.

Exercise

50 A motor is used to lift a 10 kg mass 2 m above the ground in 4 s. If the power input to the motor is 100 W, what is the efficiency of the motor?

51 A motor is 70% efficient. If 60 kJ of energy is put into the engine, how much work is got out?

52 The drag force that resists the motion of a car travelling at 80 km h^{-1} is 300 N.
- **(a)** What power is required to keep the car travelling at that speed?
- **(b)** If the efficiency of the engine is 60%, what is the power of the engine?

2.10 Uniform circular motion

Assessment statements

2.4.1 Draw a vector diagram to illustrate that the acceleration of a particle moving with constant speed in a circle is directed towards the centre of the circle.
2.4.2 Apply the expression for centripetal acceleration.
2.4.3 Identify the force producing circular motion in various situations.
2.4.4 Solve problems involving circular motion.

Circular motion quantities
When describing motion in a circle we often use quantities referring to the angular rather than linear quantities.

Time period (T)
The time taken for one complete circle.

Angular displacement (θ)
The angle swept out by a line joining the body to the centre.

Angular velocity (ω)
The angle swept out per unit time.
$\omega = \left(\frac{2\pi}{T}\right)$

Frequency (f)
The number of complete circles per unit time $\left(f = \frac{1}{T}\right)$.

Size of centripetal acceleration

The centripetal acceleration is given by the following formula

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

Or in circular terms

$$a = \omega^2 r$$

If a car travels around a bend at 30 km h^{-1}, it is obviously travelling at a constant speed, since the speedometer would register 30 km/hr all the way round. However it is not travelling at constant velocity. This is because velocity is a vector quantity and for a vector quantity to be constant, both magnitude and direction must remain the same. Bends in a road can be many different shapes, but to simplify things, we will only consider circular bends.

Centripetal acceleration

From the definition of acceleration, we know that if the velocity of a body changes, it must be accelerating, and that the direction of acceleration is in the direction of the change in velocity. Let us consider a body moving in a circle with a constant speed v. Figure 2.86 shows two positions of the body separated by a short time.

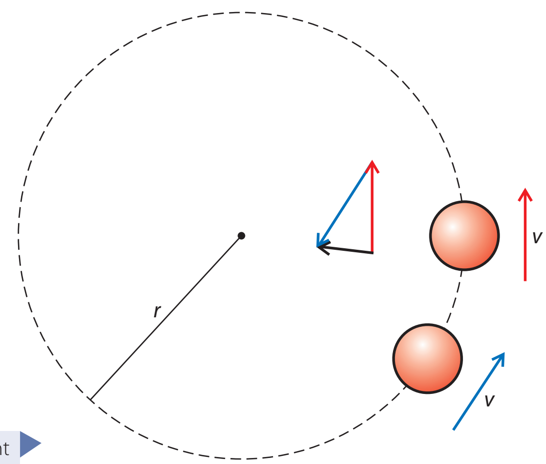


Figure 2.86 A body travels at constant speed around a circle of radius r.

Since this is a two-dimensional problem, we will need to draw the vectors to find out the change in velocity. This has been done on the diagram. It can be seen that the vector representing the change in velocity points to the centre of the circle. This means that the acceleration is also directed towards the centre, and this acceleration is called *centripetal acceleration*. All bodies moving in a circle accelerate towards the centre.

Centripetal force

From Newton's first law, we know that if a body accelerates, there must be an unbalanced force acting on it. The second law tells us that that force is in the direction of the acceleration. This implies that there must be a force acting towards the centre. This force is called the *centripetal force*.

Centripetal force, $F = \frac{mv^2}{r} = m\omega^2 r$

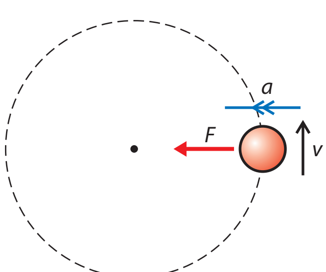


Figure 2.87 Showing the direction of the force and acceleration.

Examples

All bodies moving in a circle must be acted upon by a force towards the centre of the circle. However, this can be provided by many different forces.

1. Ball on a string

You can make a ball move in a circle by attaching it to a string and swinging it round your head. In this case the centripetal force is provided by the tension in the string. If the string suddenly broke, the ball would fly off in a straight line at a tangent to the circle.

2. Car going round a bend

When a car goes round a bend, the force causing it to move in a circle is the friction between the road and the tyres. Without this friction the car would move in a straight line.

Wall of Death

3. Wall of Death

In the Wall of Death the rider travels around the inside walls of a cylinder. Here the centripetal force is provided by the normal force between the wall and the tyres.

Exercise

53 Calculate the centripetal force for a 1000 kg car travelling around a circular track of radius 50 m at 30 km h^{-1}.

54 A 200 g ball is made to travel in a circle of radius 1 m on the end of a string. If the maximum force that the string can withstand before breaking is 50 N, what is the maximum speed of the ball?

55 A body of mass 250 g moves in a circle of radius 50 cm with a speed of 2 m s^{-1}. Calculate
- **(a)** the distance travelled per revolution
- **(b)** the time taken for one revolution
- **(c)** the angle swept out per revolution
- **(d)** the angular velocity
- **(e)** the centripetal acceleration
- **(f)** the centripetal force.

Centripetal is not an extra force
Remember when solving circular motion problems, centripetal force is not an extra force – it is one of the existing forces. Your task is to find which force (or a component of it) points towards the centre.

Practice questions

1 This question is about linear motion.
A police car P is stationary by the side of a road. A car S, exceeding the speed limit, passes the police car P at a constant speed of 18 m s^{-1}. The police car P sets off to catch car S just as car S passes the police car P. Car P accelerates at 4.5 m s^{-2} for a time of 6.0 s and then continues at constant speed. Car P takes a time *t* seconds to draw level with car S.

(a) **(i)** State an expression, in terms of *t*, for the distance car S travels in *t* seconds. (1)

(ii) Calculate the distance travelled by the police car P during the first 6.0 seconds of its motion. (1)

(iii) Calculate the speed of the police car P after it has completed its acceleration. (1)

(iv) State an expression, in terms of t, for the distance travelled by the police car P during the time that it is travelling at constant speed. (1)

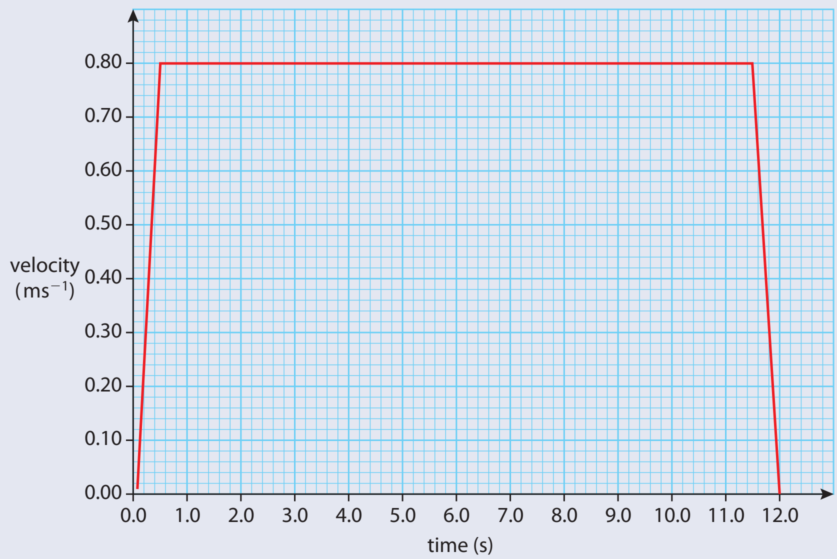
(b) Using your answers to **(a)**, determine the total time *t* taken for the police car P to draw level with car S. (2)

(Total 6 marks)

2 This question is about the kinematics of an elevator (lift).

(a) Explain the difference between the gravitational mass and the inertial mass of an object. (3)

An elevator (lift) starts from rest on the ground floor and comes to rest at a higher floor. Its motion is controlled by an electric motor. A simplified graph of the variation of the elevator's velocity with time is shown below.



(b) The mass of the elevator is 250 kg. Use this information to calculate

(i) the acceleration of the elevator during the first 0.50 s. (2)

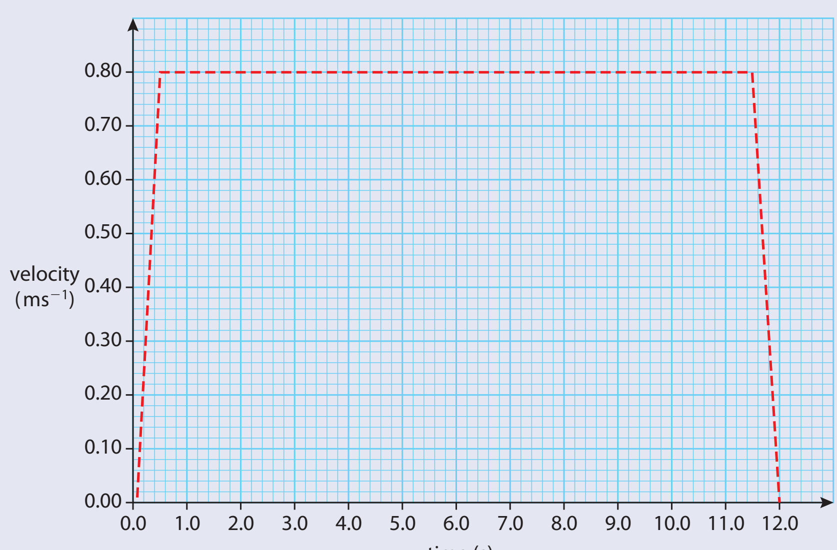
(ii) the total distance travelled by the elevator. (2)

(iii) the minimum work required to raise the elevator to the higher floor. (2)

(iv) the minimum average power required to raise the elevator to the higher floor. (2)

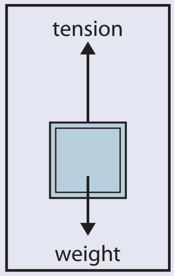
(v) the efficiency of the electric motor that lifts the elevator, given that the input power to the motor is 5.0 kW. (2)

(c) On the graph axes below, sketch a realistic variation of velocity for the elevator. Explain your reasoning. *(The simplified version is shown as a dotted line)*



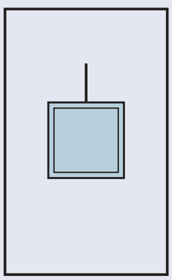
(2)

The elevator is supported by a cable. The diagram opposite is a free-body force diagram for when the elevator is moving upwards during the first 0.50 s.

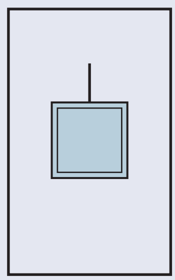


(d) In the space below, draw free-body force diagrams for the elevator during the following time intervals.

(i) 0.5 to 11.50 s



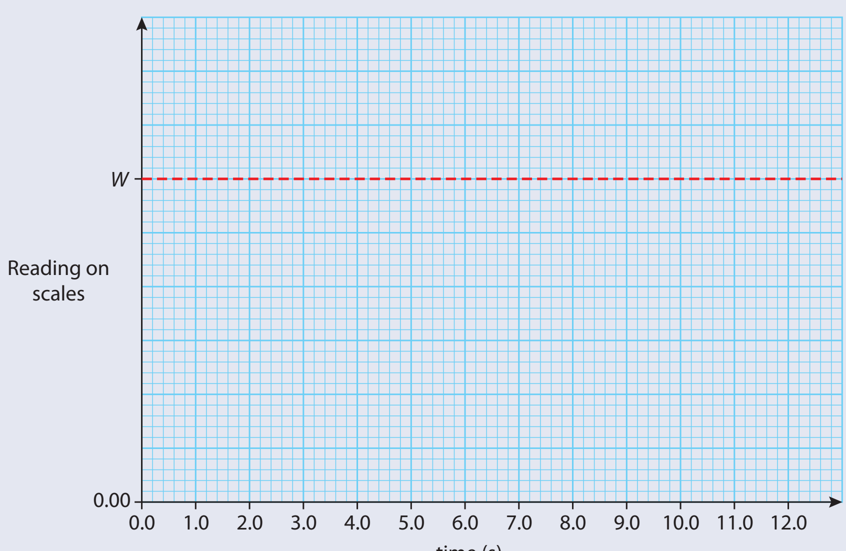
(ii) 11.50 to 12.00 s



(3)

A person is standing on weighing scales in the elevator. Before the elevator rises, the reading on the scales is *W*.

(e) On the axes below, sketch a graph to show how the reading on the scales varies during the whole 12.00 s upward journey of the elevator. *(Note that this is a sketch graph – you do not need to add any values.)*



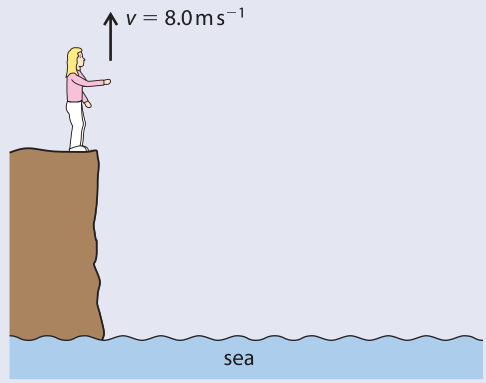
(3)

(f) The elevator now returns to the ground floor where it comes to rest. Describe and explain the energy changes that take place during the whole up and down journey.

(4)

(Total 25 marks)

3 This question is about throwing a stone from a cliff.
Antonia stands at the edge of a vertical cliff and throws a stone vertically upwards.



The stone leaves Antonia's hand with a speed $v = 8.0\,\text{m s}^{-1}$.
The acceleration of free fall g is $10\,\text{m s}^{-2}$ and all distance measurements are taken from the point where the stone leaves Antonia's hand.

(a) Ignoring air resistance calculate

(i) the maximum height reached by the stone. (2)

(ii) the time taken by the stone to reach its maximum height. (1)

The time between the stone leaving Antonia's hand and hitting the sea is 3.0 s.

(b) Determine the height of the cliff. (3)

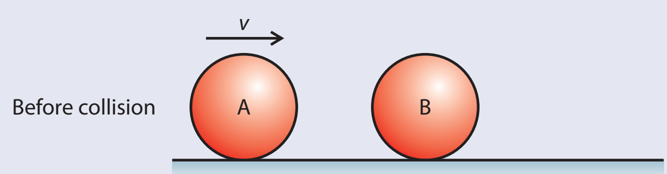
(Total 6 marks)

4 This question is about conservation of momentum and conservation of energy.

(a) State Newton's third law. (1)

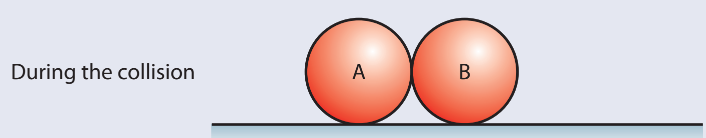
(b) State the law of conservation of momentum. (2)

The diagram below shows two identical balls A and B on a horizontal surface. Ball B is at rest and ball A is moving with speed v along a line joining the centres of the balls. The mass of each ball is M.



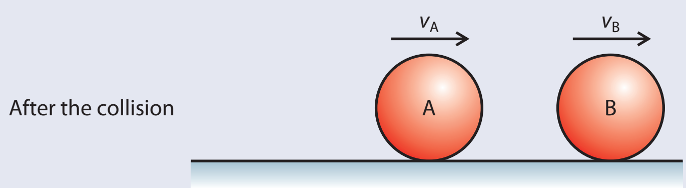
During the collision of the balls, the magnitude of the force that ball A exerts on ball B is F_{AB} and the magnitude of the force that ball B exerts on ball A is F_{BA}.

(c) On the diagram below, add labelled arrows to represent the magnitude and direction of the forces F_{AB} and F_{BA}.



(3)

The balls are in contact for a time t. After the collision, the speed of ball A is $+v_A$ and the speed of ball B is $+v_B$ in the directions shown.



As a result of the collision, there is a change in momentum of ball A and of ball B.

(d) Use Newton's second law of motion to deduce an expression relating the forces acting during the collision to the change in momentum of

(i) ball B. (2)

(ii) ball A. (2)

(e) Apply Newton's third law and your answers to **(d)** to deduce that the change in momentum of the system (ball A and ball B) as a result of this collision, is zero. (4)

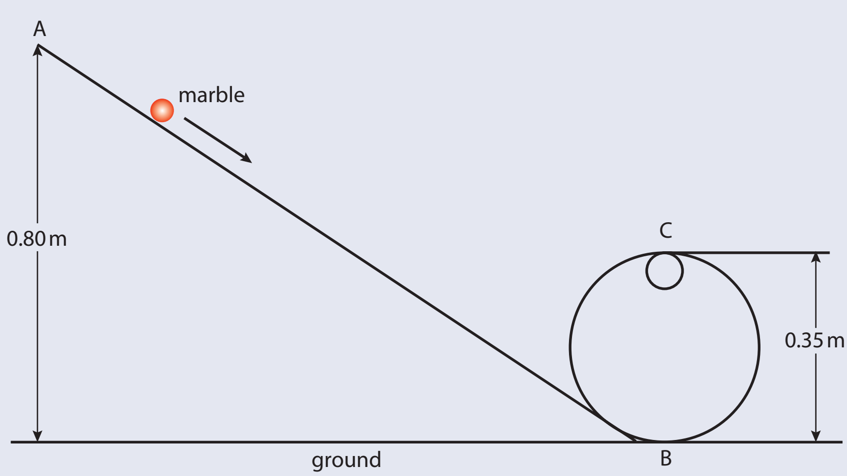
(f) Deduce that if kinetic energy is conserved in the collision, then after the collision, ball A will come to rest and ball B will move with speed v. (3)

(Total 17 marks)

5 This question is about the kinematics and dynamics of circular motion.

(a) A car goes round a curve in a road at constant speed. Explain why, although its speed is constant, it is accelerating. (2)

In the diagram below, a marble (small glass sphere) rolls down a track, the bottom part of which has been bent into a loop. The end A of the track, from which the marble is released, is at a height of 0.80 m above the ground. Point B is the lowest point and point C the highest point of the loop. The diameter of the loop is 0.35 m.



The mass of the marble is 0.050 kg. Friction forces and any gain in kinetic energy due to the rotating of the marble can be ignored. The acceleration due to gravity, $g = 10\ \text{m s}^{-2}$.

Consider the marble when it is at point C.

(b) **(i)** On the diagram, draw an arrow to show the direction of the resultant force acting on the marble. (1)

(ii) State the names of the two forces acting on the marble. (2)

(iii) Deduce that the speed of the marble is $3.0\ \text{m s}^{-1}$. (3)

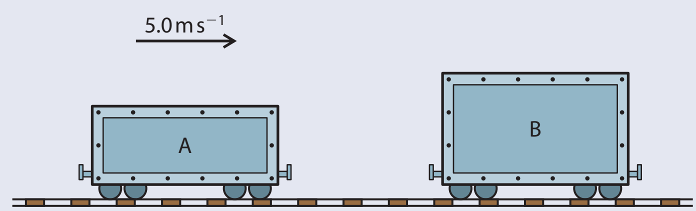
(iv) Determine the resultant force acting on the marble and hence determine the reaction force of the track on the marble. (4)

(Total 12 marks)

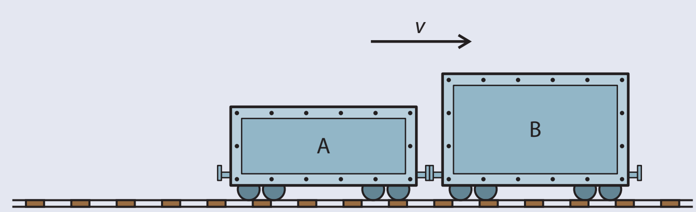
6 This question is about the collision between two railway trucks (carts).

(a) Define *linear momentum*. (1)

In the diagram below, railway truck A is moving along a horizontal track. It collides with a stationary truck B and on collision, the two join together. Immediately before the collision, truck A is moving with speed 5.0 m s^{-1}. Immediately after collision, the speed of the trucks is v.



Immediately before collision



Immediately after collision

The mass of truck A is 800 kg and the mass of truck B is 1200 kg.

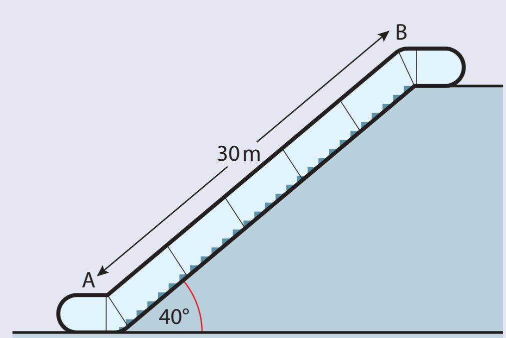
(b) **(i)** Calculate the speed v immediately after the collision. (3)

(ii) Calculate the total kinetic energy lost during the collision. (2)

(c) Suggest what has happened to the lost kinetic energy. (2)

(Total 8 marks)

7 This question is about estimating the energy changes for an escalator (moving staircase). The diagram below represents an escalator. People step on to it at point A and step off at point B.



(a) The escalator is 30 m long and makes an angle of 40° with the horizontal. At full capacity, 48 people step on at point A and step off at point B every minute.

(i) Calculate the potential energy gained by a person of weight 700 N in moving from A to B. (2)

(ii) Estimate the energy supplied by the escalator motor to the people every minute when the escalator is working at full capacity. (1)

(iii) State **one** assumption that you have made to obtain your answer to **(ii)**. (1)

The escalator is driven by an electric motor that has an efficiency of 70 %.

(b) **(i)** Using your answer to **(a)(ii),** calculate the minimum input power required by the motor to drive the escalator. (3)

(ii) Explain why it is not necessary to take into account the weight of the escalator when calculating the input power. (1)

(c) Explain why in practice, the power of the motor will need to be greater than that calculated in **(b)(i)**. (1)

(Total 9 marks)

3 Thermal physics

3.1 Thermal concepts

Assessment statements

3.1.1 State that temperature determines the direction of thermal energy transfer between two objects.
3.1.2 State the relation between the Kelvin and Celsius scales of temperature.
3.1.3 State that the internal energy of a substance is the total potential energy and random kinetic energy of the molecules of the substance.
3.1.4 Explain and distinguish between the macroscopic concepts of temperature, internal energy and thermal energy (heat).
3.1.5 Define the mole and molar mass.
3.1.6 Define the Avogadro constant.

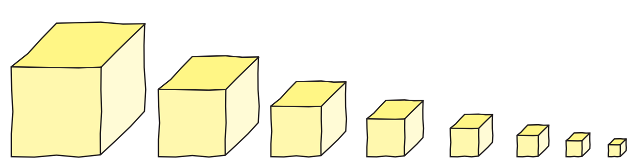
The role of the physicist is to observe our physical surroundings, take measurements and think of ways to explain what we see. Up to this point in the course we have been dealing with the motion of bodies. We can describe bodies in terms of their mass and volume, and if we know their speed and the forces that act on them, we can calculate where they will be at any given time. We even know what happens if two hit each other. However, this is not enough to describe all the differences between objects. For example, by simply holding different objects, we can feel that some are hot and some are cold.

In this chapter we will develop a model to explain these differences, but first of all we need to know what is inside matter.

The particle model of matter

Ancient Greek philosophers spent a lot of time thinking about what would happen if they took a piece of cheese and kept cutting it in half.

Figure 3.1 Can we keep cutting the cheese for ever?



They didn't think it was possible to keep halving it for ever, so they suggested that there must exist a smallest part – this they called the *atom*.

Atoms are too small to see (about 10^{-10} m in diameter) but we can think of them as very small perfectly elastic balls. This means that when they collide, both momentum and kinetic energy are conserved.

Elements and compounds

We might ask: 'If everything is made of atoms, why isn't everything the same?' The answer is that there are many different types of atom.

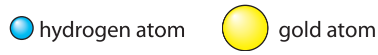


Figure 3.2 Gold is made of gold atoms and hydrogen is made of hydrogen atoms.

There are 117 different types of atom, and a material made of just one type of atom is called an *element*. There are, however, many more than 117 different types of material. The other types of matter are made of atoms that have joined together to form molecules. Materials made from molecules that contain more than one type of atom are called *compounds*.

This is a good example of how models are used in physics. Here we are modelling something that we can't see, the atom, using a familiar object, a rubber ball.

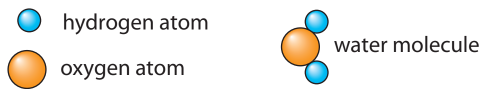


Figure 3.3 Water is an example of a compound.

The mole

When buying apples, you can ask for 5 kg of apples, or, say, 10 apples – both are a measure of amount. It's the same with matter – you can express amount in terms of either mass or number of particles.

A mole of any material contains 6.022×10^{23} atoms or molecules; this number is known as *Avogadro's number*.

Although all moles have the same number of particles, they don't all have the same mass. A mole of carbon has a mass of 12 g and a mole of neon has a mass of 20 g – this is because a neon atom has more mass than a carbon atom.

Moles of different compounds.

The three states of matter

From observations we know that there are three types, or *states* of matter: *solid*, *liquid* and *gas*. If the particle model is correct, then we can use it to explain why the three states are different.

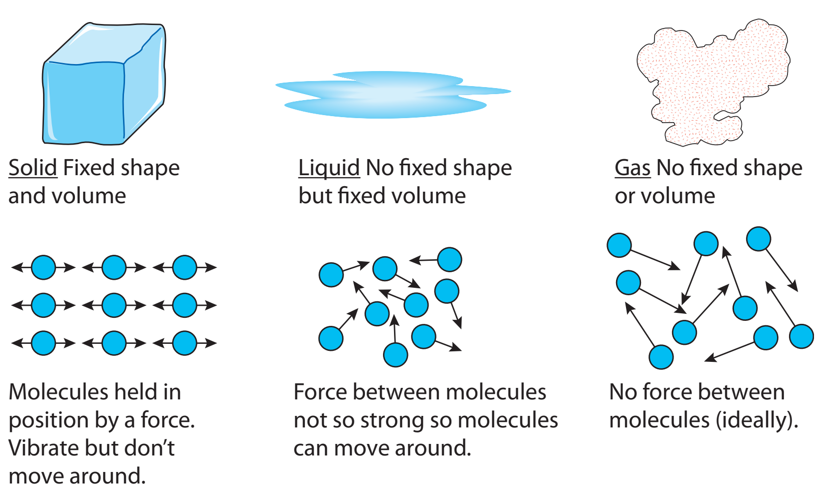


Figure 3.4 The particle model explains the differences between solids, liquids and gases. (The arrows represent velocity vectors.)

We can't prove that this model is true – we can only provide evidence that supports it.

Ice, water and steam.

Worked example

1 If a mole of carbon has a mass of 12 g, how many atoms of carbon are there in 2 g?

2 The density of iron is 7874 kg m^{-3} and the mass of a mole of iron is 55.85 g. What is the volume of 1 mole of iron?

Solution

1 One mole contains 6.022×10^{23} atoms.

2 g is $\frac{1}{6}$ of a mole so contains $\frac{1}{6} \times 6.022 \times 10^{23}$ atoms $=$ **1.004×10^{23} atoms**

2
$$\text{density} = \frac{\text{mass}}{\text{volume}}$$
$$\text{volume} = \frac{\text{mass}}{\text{density}}$$
$$\text{Volume of 1 mole} = \frac{0.05585}{7874}\,\text{m}^3$$
$$= 7.093 \times 10^{-6}\,\text{m}^3$$
$$= \mathbf{7.09\,cm^3}$$

- **Examiner's hint:** Be careful with the units. Do all calculations using m^3.

Exercises

1 The mass of 1 mole of copper is 63.54 g and its density 8920 kg m^{-3}
  (a) What is the volume of one mole of copper?
  (b) How many atoms does one mole of copper contain?
  (c) How much volume does one atom of copper occupy?

2 If the density of aluminium is 2700 kg m^{-3} and the volume of 1 mole is 10 cm^3, what is the mass of one mole of aluminium?

Internal energy

In Chapter 2, Mechanics, you met the concepts of energy and work. Use these concepts to consider the following:

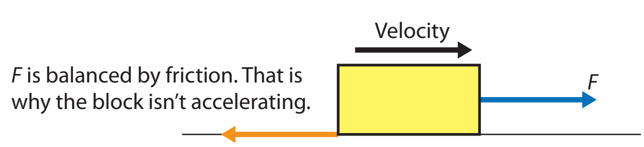


Figure 3.5 A wooden block is pulled along a rough horizontal surface at a constant velocity.

Is any work being done on the block by force *F*?

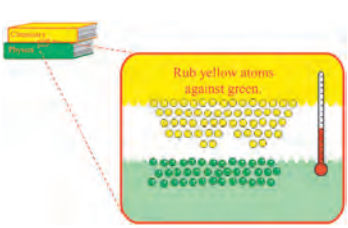
Is energy being transferred to the block?

Is the KE of the block increasing?

Is the PE of the block increasing?

Where is the energy going?

You will have realised that since work is done, energy is given to the block, but its PE and KE are not increasing. Since energy is conserved, the energy must be going somewhere. It is going inside the block as internal energy. We can explain what is happening using the particle model.



To a view a simulation showing how friction can increase temperature, visit heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 3.1.

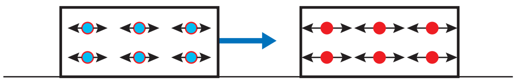


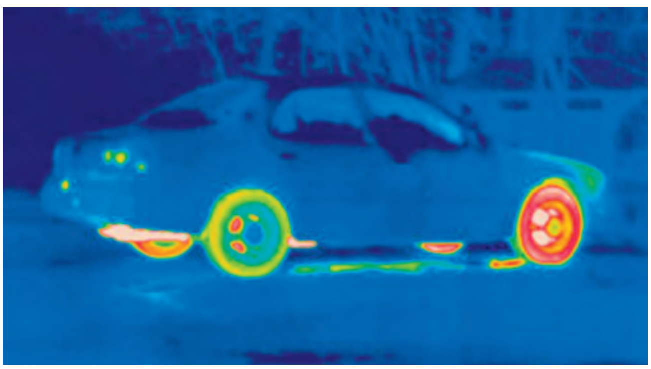
Figure 3.6 Molecules gain internal energy.

When we do work on an object, it enables the molecules to move faster (increasing KE) and move apart (increasing PE). We say that the *internal energy* of the object has increased.

In a solid, this means increasing the KE and PE of the molecules; in a gas it is just the KE. This is because there are no forces between the molecules of a gas, so it doesn't require any work to pull them apart.

Worked example

1 A car of mass 1000 kg is travelling at 30 m s^{-1}. If the brakes are applied, how much heat energy is transferred to the brakes?



This thermogram of a car shows how the wheels have become hot owing to friction between the road and the tyres, and the brakes pads and discs.

Solution

When the car is moving it has kinetic energy. This must be transferred to the brakes when the car stops.

$$\begin{aligned} \text{KE} &= \tfrac{1}{2}mv^2 \\ &= \tfrac{1}{2} \times 1000 \times 30^2\ \text{J} \\ &= 450\ \text{kJ} \end{aligned}$$

So thermal energy transferred to the brakes = 450 kJ

Exercises

3 A block of metal, mass 10 kg, is dropped from a height of 40 m.
  (a) How much energy does the block have before it is dropped?
  (b) How much heat energy do the block and floor gain when it hits the floor?

4 If the car in Example 1 was travelling at 60 m s^{-1}, how much heat energy would the brakes receive?

Temperature

If we now pick a block up, after dragging it, we will notice something has changed. It has got hot; doing work on the block has made it hot. *Hotness* and *coldness* are the ways we perceive differences between objects. In physics, we use temperature to measure this difference more precisely.

It is important to realise the difference between perception and physical measurement.

Temperature (T) is a measure of how hot or cold an object is, and it is temperature that determines the direction of heat flow.

Temperature is a *scalar* quantity, and is measured in degrees Celsius (°C) or kelvin (K).

0 °C is equivalent to 273 K.
100 °C is equivalent to 373 K.

At normal atmospheric pressure, pure water boils at 100 °C and freezes at 0 °C. Room temperature is about 20 °C.

During the first part of this chapter, we will measure temperature in Celsius. However when dealing with gases, we will use kelvin – this is because the Kelvin scale is based on the properties of a gas.

To convert from degrees Celsius to kelvin, simply add 273.

Thermometers

Temperature cannot be measured directly, so we have to find something that changes when the temperature changes. The most common thermometer consists of a small amount of alcohol in a thin glass tube. As temperature increases, the volume of the alcohol increases, so it rises up the tube. When we measure temperature, we are really measuring the length of the alcohol column, but the scale is calibrated to give the temperature in °C.

Temperature and the particle model

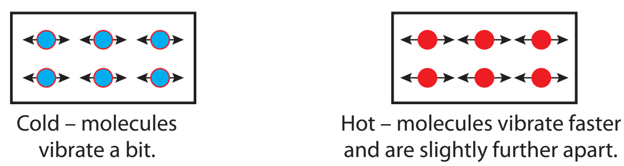


Figure 3.7 Temperature is related to kinetic energy.

From the previous model, we can see that the particles in a hot body move faster than those in a cold one. The temperature is related to the average KE of the particles.

Heat transfer

Pulling a block of wood along a rough surface is not the only way to increase its temperature. We can make a cold body hot by placing it next to a hot body. We know that if the cold body gets hot, then it must have received energy – this is heat or *thermal* energy.

We are often more interested in preventing heat flow than causing it. Placing an insulating layer (e.g. woollen cloth) between the hot and cold bodies will reduce the rate of heat flow.

Thermal equilibrium

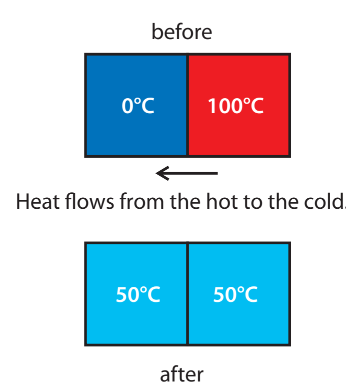


Figure 3.8 Heat flows from the hot body to the cold body until they are at the same temperature.

At this point no more heat will flow – this is called *thermal equilibrium*.

3.2 Thermal properties of matter

Assessment statements

3.2.1 Define *specific heat capacity* and *thermal capacity*.
3.2.2 Solve problems involving specific heat capacities and thermal capacities.
3.2.3 Explain the physical differences between the solid, liquid and gaseous phases in terms of molecular structure and particle motion.
3.2.4 Describe and explain the process of phase changes in terms of molecular behaviour.
3.2.5 Explain in terms of molecular behaviour why temperature does not change during a phase change.
3.2.6 Distinguish between evaporation and boiling.
3.2.7 Define *specific latent heat*.
3.2.8 Solve problems involving specific latent heats.

Thermal capacity (C)

If heat is added to a body, its temperature rises, but the actual increase in temperature depends on the body.

The *thermal capacity* (C) of a body is the amount of heat needed to raise its temperature by 1°C. Unit: J °C^{-1}

If the temperature of a body increases by an amount ΔT when quantity of heat Q is added, then the thermal capacity is given by the equation:

$$C = \frac{Q}{\Delta T}$$

This applies not only when things are given heat, but also when they lose heat.

Worked example

1 If the thermal capacity of a quantity of water is 5000 J °C^{-1}, how much heat is required to raise its temperature from 20 °C to 100 °C?

2 How much heat is lost from a block of metal of thermal capacity 800 J °C^{-1} when it cools down from 60 °C to 20 °C?

Solution

1	Thermal capacity	$Q = \frac{Q}{\Delta T}$	From definition
	So	$Q = \mathrm{C}\Delta \mathrm{T}$	Rearranging
	Therefore	$Q = 5000 \times (100 - 20)$ J	
	So the heat required	$Q =$ **400 kJ**	
2	Thermal capacity	$C = \frac{Q}{\Delta T}$	From definition
	So	$Q = C\Delta T$	Rearranging
	Therefore	$Q = 800 \times (60 - 20)$ J	
	So the heat lost	$Q =$ **32 kJ**	

Exercises

5 The thermal capacity of a 60 kg human is 210 kJ °C^{-1}. How much heat is lost from a body if its temperature drops by 2 °C?

6 The temperature of a room is 10 °C. In 1 hour the room is heated to 20 °C by a 1 kW electric heater.
   (a) How much heat is delivered to the room?
   (b) What is the thermal capacity of the room?
   (c) Does all this heat go to heat the room?

● **Examiner's hint:** Remember, power is energy per unit time.

Specific heat capacity (*c*)

The thermal capacity depends on the size of the object and what it is made of. The *specific heat capacity* depends only on the material. Raising the temperature of 1 kg of water requires more heat than raising 1kg of steel by the same amount, so the specific heat capacity of water is higher than that of steel.

The specific heat capacity of a material is the amount of heat required to raise the temperature of 1kg of the material by 1°C. Unit: J kg^{-1} °C^{-1}

If a quantity of heat Q is required to raise the temperature of a mass m of material by ΔT then the specific heat capacity (c) of that material is given by the following equation:

$$c = \frac{Q}{m\Delta T}$$

The specific heat capacity of water is quite high, so it takes a lot of energy to heat up the water for a shower.

Worked example

1 The specific heat capacity of water is 4200 J kg^{-1} °C^{-1}. How much heat will be required to heat 300 g of water from 20 °C to 60 °C?

2 A metal block of mass 1.5 kg loses 20 kJ of heat. As this happens, its temperature drops from 60 °C to 45 °C. What is the specific heat capacity of the metal?

Solution

1 Specific heat capacity $c = \frac{Q}{m\Delta T}$ From definition

So $Q = cm\Delta T$ Rearranging

Therefore $Q = 4200 \times 0.3 \times 40$ Note: Convert g to kg

$Q = \mathbf{50.4\ kJ}$

2 Specific heat capacity $c = \frac{Q}{m\Delta T}$ From definition

So $c = 20\,000/1.5(60-45)$ Rearranging

$c = \mathbf{888.9\ J\,kg^{-1}\,°C^{-1}}$

Exercises

Use the data in the table to solve the problems:

Substance	Specific heat capacity (J kg^{-1} °C^{-1})
Water	4200
Copper	380
Aluminium	900
Steel	440

7 How much heat is required to raise the temperature of 250 g of copper from 20 °C to 160 °C?

8 The density of water is 1000 kg m^{-3}.
- **(a)** What is the mass of 1 litre of water?
- **(b)** How much energy will it take to raise the temperature of 1litre of water from 20 °C to 100 °C?
- **(c)** A water heater has a power rating of 1 kW. How many seconds will this heater take to boil 1 litre of water?

9 A 500 g piece of aluminium is heated with a 500 W heater for 10 minutes.
- **(a)** How much energy will be given to the aluminium in this time?
- **(b)** If the temperature of the aluminium was 20°C at the beginning, what will its temperature be after 10 minutes?

10 A car of mass 1500 kg travelling at 20 m s^{-1} brakes suddenly and comes to a stop.
- **(a)** How much KE does the car lose?
- **(b)** If 75% of the energy is given to the front brakes, how much energy will they receive?
- **(c)** The brakes are made out of steel and have a total mass of 10 kg. By how much will their temperature rise?

11 The water comes out of a showerhead at a temperature of 50 °C at a rate of 8 litres per minute.
- **(a)** If you take a shower lasting 10 minutes, how many kg of water have you used?
- **(b)** If the water must be heated from 10 °C, how much energy is needed to heat the water?

Change of state

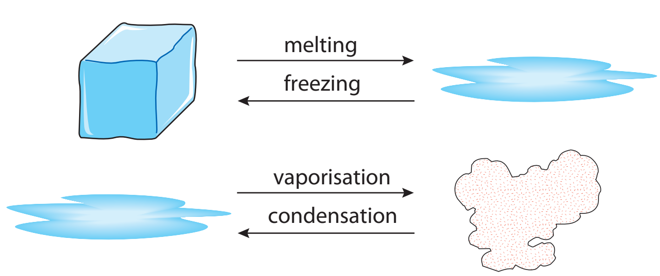
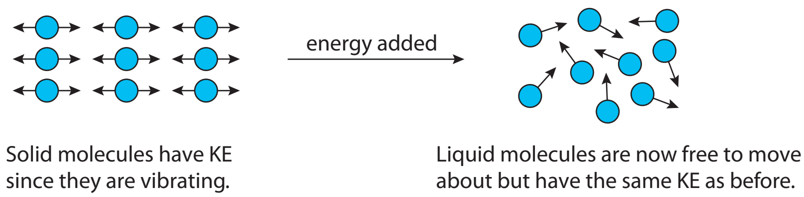


Figure 3.9 When matter changes from liquid to gas, or solid to liquid, it is changing state.

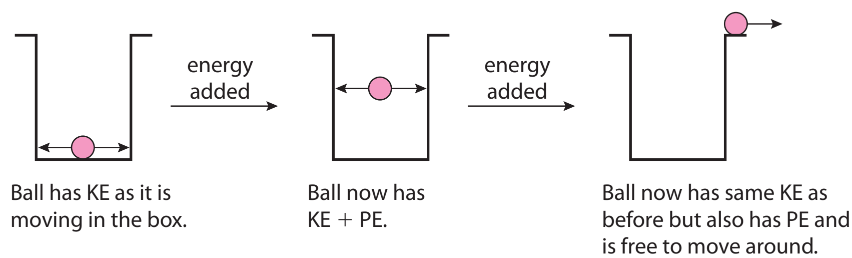
When water boils, this is called a *change of state* (or *change of phase*). As this happens, the temperature of the water doesn't change – it stays at 100 °C. In fact, we find that whenever the state of a material changes, the temperature stays the same. We can explain this in terms of the particle model.

Figure 3.10 Molecules gain PE when the state changes.



When matter changes state, the energy is needed to enable the molecules to move more freely. To understand this, consider the example below.

Figure 3.11 A ball-in-a-box model of change of state.



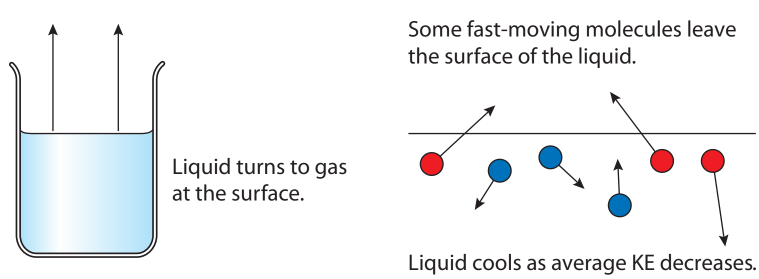
An iceberg melts as it floats into warmer water.

Boiling and evaporation

These are two different processes by which liquids can change to gases.

Boiling takes place throughout the liquid and always at the same temperature. *Evaporation* takes place only at the surface of the liquid and can happen at all temperatures.

Figure 3.12 A microscopic model of evaporation.



When a liquid evaporates, the fastest-moving particles leave the surface. This means that the average kinetic energy of the remaining particles is lower, resulting in a drop in temperature.

The rate of evaporation can be increased by:

- Increasing the surface area; this increases the number of molecules near the surface, giving more of them a chance to escape.
- Blowing across the surface. After molecules have left the surface they form a small 'vapour cloud' above the liquid. If this is blown away, it allows further molecules to leave the surface more easily.
- Raising the temperature; this increases the kinetic energy of the liquid molecules, enabling more to escape.

People sweat to increase the rate at which they lose heat. When you get hot, sweat comes out of your skin onto the surface of your body. When the sweat evaporates, it cools you down. In a sauna there is so much water vapour in the air that the sweat doesn't evaporate.

Specific latent heat (*L*)

The *specific latent heat* of a material is the amount of heat required to change the state of 1kg of the material without change of temperature.

Unit: $\mathrm{J\,kg^{-1}}$

Latent means *hidden*. This name is used because when matter changes state, the heat added does not cause the temperature to rise, but seems to disappear.

If it takes an amount of energy Q to change the state of a mass m of a substance, then the specific latent heat of that substance is given by the equation:

$$L = \frac{Q}{m}$$

Solid→liquid
Specific latent heat of fusion

Liquid→gas
Specific latent heat of vaporization

Worked example

1 The specific latent heat of fusion of water is $3.35 \times 10^5\ \mathrm{J\,kg^{-1}}$. How much energy is required to change 500 g of ice into water?

2 The amount of heat released when 100 g of steam turns to water is 2.27×10^5 J. What is the specific latent heat of vaporization of water?

Solution

1 The latent heat of fusion $L_f = \frac{Q}{m}$ From definition

So $Q = mL$ Rearranging

Therefore $Q = 0.5 \times 3.35 \times 10^5$ J

So the heat required $Q = \mathbf{1.675 \times 10^5 J}$

2 The specific latent heat of vaporization $L = \frac{Q}{m}$ From definition

Therefore $L = 2.27 \times 10^5/0.1\ \mathrm{J\,kg^{-1}}$

So the specific latent heat of vaporization $L = \mathbf{2.27 \times 10^6\ J\,kg^{-1}}$

This equation $\left(L = \frac{Q}{m}\right)$ can also be used to calculate the heat lost when a substance changes from gas to liquid, or liquid to solid.

Exercises

Latent heats of water

Latent heat of vaporization	$2.27 \times 10^6\ \mathrm{J\,kg^{-1}}$
Latent heat of fusion	$3.35 \times 10^5\ \mathrm{J\,kg^{-1}}$

Use the data about water in the table to solve the following problems.

12 If the mass of water in a cloud is 1million kg, how much energy will be released if the cloud turns from water to ice?

13 A water boiler has a power rating of 800 W. How long will it take to turn 400 g of boiling water into steam?

14 The ice covering a 1000 $\mathrm{m^2}$ lake is 2 cm thick.

(a) If the density of ice is 920 $\mathrm{kg\,m^{-3}}$, what is the mass of the ice on the lake?
(b) How much energy is required to melt the ice?
(c) If the sun melts the ice in 5 hours, what is the power delivered to the lake?
(d) How much power does the Sun deliver per $\mathrm{m^2}$?

Graphical representation of heating

The increase of the temperature of a body can be represented by a temperature–time graph. Observing this graph can give us a lot of information about the heating process.

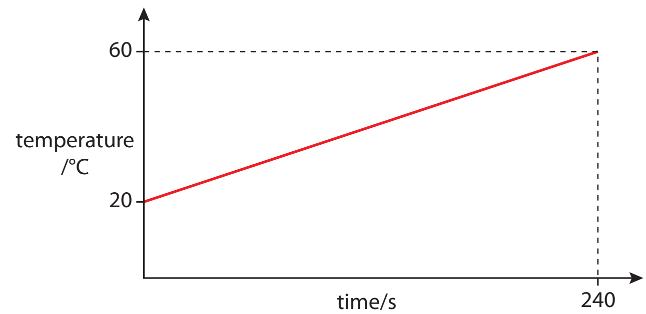


Figure 3.13 Temperature–time graph for 1kg of water being heated in an electric kettle.

In this example, we are ignoring the heat given to the kettle and the heat lost.

From this graph we can calculate the amount of heat given to the water per unit time (power).

The gradient of the graph $= \dfrac{\text{temperature rise}}{\text{time}} = \dfrac{\Delta T}{t}$

We know from the definition of specific heat capacity that

$$\text{heat added} = mc\Delta T$$

The rate of adding heat $= P = \dfrac{mc\Delta T}{t}$

So $P = mc \times \text{gradient}$

The gradient of this line $= \dfrac{(60 - 20)}{240}\,°\mathrm{C\,s^{-1}} = 0.167\,°\mathrm{C\,s^{-1}}$

So the power delivered $= 4200 \times 0.167\,\mathrm{W} = 700\,\mathrm{W}$

If we continue to heat this water it will begin to boil.

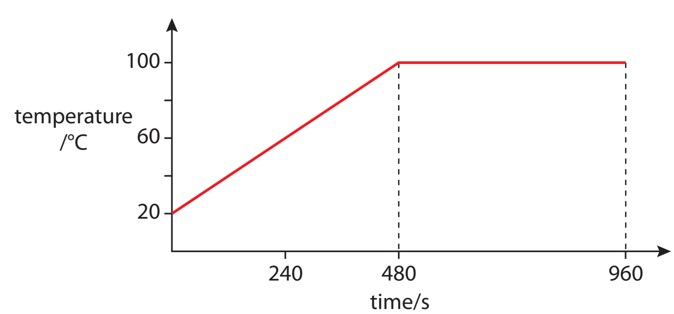


Figure 3.14 A graph of temperature vs time for boiling water. When the water is boiling, the temperature does not increase any more.

If we assume that the heater is giving heat to the water at the same rate, then we can calculate how much heat was given to the water whilst it was boiling.

Power of the heater $= 700\,\mathrm{W}$

Time of boiling $= 480\,\mathrm{s}$

Energy supplied $= \text{power} \times \text{time} = 700 \times 480\,\mathrm{J} = 3.36 \times 10^5\,\mathrm{J}$

From this we can calculate how much water must have turned to steam.

Heat added to change state $= \text{mass} \times \text{latent heat of vaporization}$,

where latent heat of vaporization of water $= 2.27 \times 10^6\,\mathrm{J\,kg^{-1}}$.

Mass changed to steam $= \dfrac{3.36 \times 10^5}{2.27 \times 10^6} = 0.15\,\mathrm{kg}$

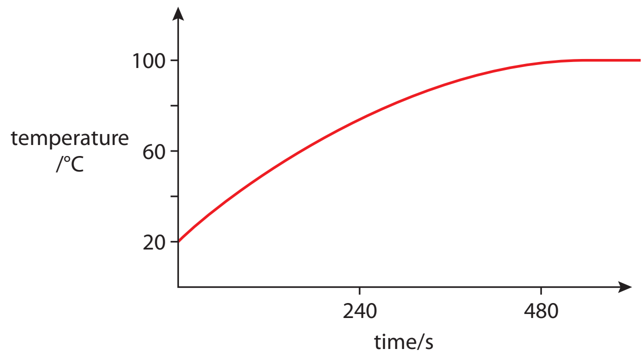


Figure 3.15 Heat loss.

When boiling a kettle, heat is continually being lost to the room. The amount of heat loss is proportional to the temperature of the kettle. For this reason, a graph of temperature against time is actually a curve, as shown in Figure 3.15.

The fact that the gradient decreases, tells us that the amount of heat given to the water gets less with time. This is because as it gets hotter, more and more of the heat is lost to the room.

Measuring thermal quantities by the method of mixtures

The method of mixtures can be used to measure the specific heat capacity and specific latent heat of substances.

Specific heat capacity of a metal

A metal sample is first heated to a known temperature. The most convenient way of doing this is to place it in boiling water for a few minutes; after this time it will be at 100 °C. The hot metal is then quickly moved to an insulated cup containing a known mass of cold water. The hot metal will cause the temperature of the cold water to rise; the rise in temperature is measured with a thermometer. Some example temperatures and masses are given in Figure 3.16.

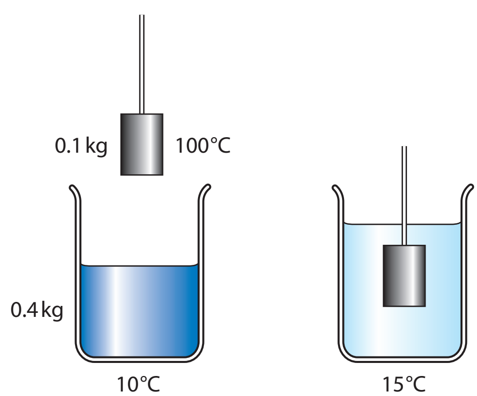


Figure 3.16 Measuring the specific heat capacity of a metal.

As the specific heat capacity of water is 4180 J kg^{-1} °C^{-1}, we can calculate the specific heat capacity of the metal.

ΔT for the metal $= 100 - 15 = 85$ °C

and ΔT for the water $= 15 - 10 = 5$ °C

Applying the formula $Q = mc\Delta T$ we get

$(mc\Delta T)_{\text{metal}} = 0.1 \times c \times 85 = 8.5c$

$(mc\Delta T)_{\text{water}} = 0.4 \times 4180 \times 5 = 8360$

If no heat is lost, then the heat transferred from the metal = heat transferred to the water

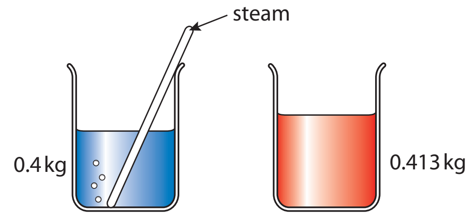
$8.5c = 8360$

$c_{\text{metal}} = 983$ J kg^{-1} °C^{-1}

Latent heat of vaporization of water

To measure the latent heat of vaporization, steam is passed into cold water. Some of the steam condenses in the water, causing the water temperature to rise. The heat from the steam = the heat to the water.

Figure 3.17 By measuring the rise in temperature, the specific latent heat can be calculated.



In Figure 3.17, 13 g of steam have condensed in the water, raising its temperature by 20 °C. The steam condenses then cools down from 100 °C to 30 °C.

Heat from steam $= ml_{\text{steam}} + mc\Delta T_{\text{water}}$

$0.013 \times L + 0.013 \times 4.18 \times 10^3 \times 70 = 0.013L + 3803.8$

Heat transferred to cold water $= mc\Delta T_{\text{water}} = 0.4 \times 4.18 \times 10^3 \times 20$

$= 33\,440\ \text{J}$

Since heat from steam = heat to water

$0.013L + 3803.8 = 33\,440$

So $L = \dfrac{33\,440 - 3803.8}{0.013}$

$L = 2.28 \times 10^6\ \text{J kg}^{-1}$

Heat loss

In both of these experiments, some of the heat coming from the hot source can be lost to the surroundings. To reduce heat loss, the temperatures can be adjusted, so you could start the experiment below room temperature and end the same amount above (e.g. if room temperature is 20 °C, then you can start at 10 °C and end at 30 °C).

Transfer of water

In the specific heat capacity experiment, droplets of hot water may be transferred with the metal block. This would add extra energy to the water, causing the temperature to rise a little bit too high. In the latent heat experiment, droplets of water sometimes condense in the tube – since they have already condensed, they don't give so much heat to the water.

3.3 Kinetic model of an ideal gas

Assessment statements

3.2.9 Define *pressure*.
3.2.10 State the assumptions of the kinetic model of an ideal gas.
3.2.11 State that temperature is a measure of the average random kinetic energy of the molecules of an ideal gas.
3.2.12 Explain the macroscopic behaviour of an ideal gas in terms of a molecular model.

The ideal gas

Of the three states of matter, the gaseous state has the simplest model; this is because the forces between the molecules of a gas are very small, so they are able to move freely. We can therefore use what we know about the motion of particles learnt in the mechanics section to study gases in more detail.

According to our simple model, a gas is made up of a large number of perfectly elastic, tiny spheres moving in random motion.

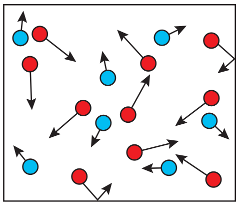


Figure 3.18 Simple model of a gas in a box. In reality the molecules have a range of velocities, not just two.

This model makes some assumptions:

- The molecules are perfectly elastic.
- The molecules are spheres.
- The molecules are identical.
- There are no forces between the molecules (except when they collide) – this means that the molecules move with constant velocity between collisions.
- The molecules are very small, that is, their total volume is much smaller than the volume of the gas.

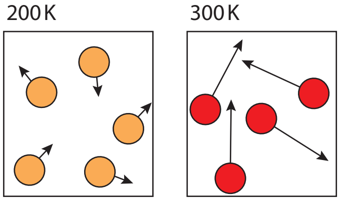
Some of these assumptions are not true for all gases, especially when the gas is compressed (when the molecules are so close together that they experience a force between them). The gas then behaves as a liquid. However, to keep things simple, we will only consider gases that behave like our model. We call these gases ideal gases.

Nitrogen becomes a liquid at low temperatures.

Temperature of a gas

From our general particle model of matter, we know that the temperature of a gas is directly related to the average KE of the molecules. If the temperature increases, then the speed of the particles will increase.

Figure 3.19 The molecules in a hot gas have a higher average KE.



Pressure of a gas

Let us apply what we know about particles to one molecule of a gas. Consider a single gas molecule in a box. According to the model, this is like a perfectly elastic sphere bouncing off the sides

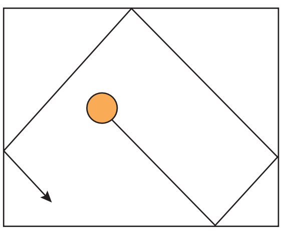


Figure 3.20 A rubber ball bouncing around a box.

We can see that this particle keeps hitting the walls of the container. Each time it does this, its direction, and therefore its velocity, changes.

Newton's first law of motion says that if a particle isn't at rest or moving with a constant velocity then it must be experiencing an unbalanced force. The particle is therefore experiencing an unbalanced force.

Newton's second law says that the size of this force is equal to the rate of change of momentum, so the force will be greater if the particle travels with a greater speed, or hits the sides more often.

Newton's third law says that if body A exerts a force on body B, then body B will exert an equal and opposite force on A. The wall exerts a force on the particle, so the particle must exert a force on the wall.

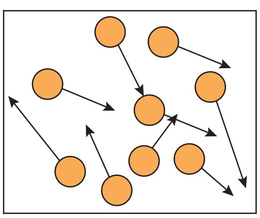


Figure 3.21 Many rubber balls bouncing around a box.

If we now add more molecules (as in Figure 3.21) then the particles exert a continuous force, F, on the walls of the container. If the walls have a total area A, then since

$$\text{pressure} = \frac{\text{force}}{\text{area}}$$

we can say that the pressure exerted on the walls is F/A — in other words, the particles exert a pressure on the container.

It is important to realise that we have been talking about the gas model, not the actual gas. The model predicts that the gas should exert a pressure on the walls of its container and it does.

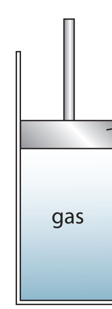
The atmosphere also exerts a pressure; this changes from day to day but is approximately 100 kPa.

1 pascal = 1 Pa = 1 $\mathrm{N\,m^{-2}}$

Properties of a gas

We can now use the particle model to explain why a gas behaves as it does.

Figure 3.22 A gas in a piston can be used to vary the properties of a gas.



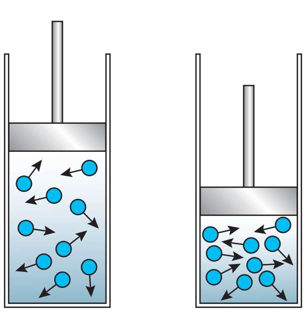


Figure 3.23 The volume of a gas is reduced.

Pressure and volume

If the volume is reduced, the particles hit the walls more often, since the walls are closer together. The force exerted by the molecules is equal to the rate of change of momentum; this will increase if the hits are more frequent, resulting in an increased pressure.

Pressure and temperature

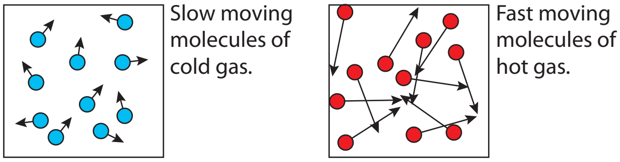
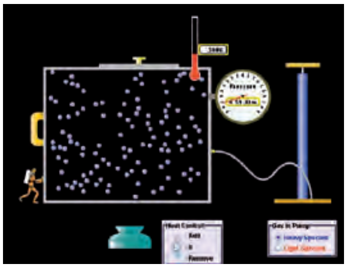


Figure 3.24 The temperature of a gas is increased.



Increase in temperature increases the speed of the molecules. When the molecules hit the walls, their change of momentum will be greater and they will hit the walls more often. The result is a greater rate of change of momentum and hence a larger force. This results in an increase in pressure.

Doing work on a gas

When you push the piston of a pump, it collides with the molecules, giving them energy (rather like a tennis racket hitting a ball). You are doing work on the gas. The increase in kinetic energy results in an increase in temperature and pressure. This is why the temperature of a bicycle pump increases when you pump up the tyres.

W To understand how pressure, temperature and volume of a gas are related, visit heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 3.2.

Gas does work

When a gas expands, it has to push away the surrounding air. In pushing the air away, the gas does work, and doing this work requires energy. This energy comes from the kinetic energy of the molecules, resulting in a reduction in temperature. This is why an aerosol feels cold when you spray it; the gas expands as it comes out of the canister.

3.4 Thermodynamics

Assessment statements

10.1.3 Describe the concept of the absolute zero of temperature and the Kelvin scale of temperature
10.1.1 State the equation of state for an ideal gas.
10.1.4 Solve problems using the equation of state of an ideal gas.
10.1.2 Describe the difference between an ideal gas and a real gas.

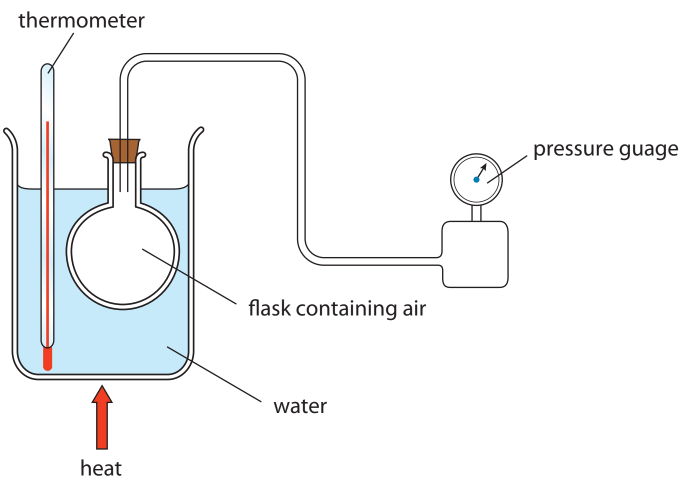


Figure 3.25 The pressure of a fixed volume of gas can be measured as the temperature is changed.

The absolute temperature scale (Kelvin)

Absolute zero

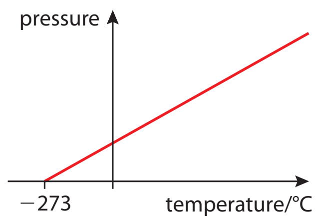


Figure 3.26

If we measure the pressure of a fixed volume of gas at different temperatures we find that temperature and pressure are linearly related as shown in Figure 3.26. We can explain this using the kinetic theory in the following way. Increasing the temperature results in an increase in the average KE of the molecules, so the molecules start to move faster. When the fast moving molecules collide with the walls of the container, the change in momentum is greater so the force exerted on the wall is greater (according to Newton). This, coupled with the fact that the collisions will now occur more often, results in an increase in pressure.

We can see from Figure 3.26 that the line passes through the x-axis at −273 °C. This is the temperature when the pressure of the gas is zero. According to the kinetic theory, this is the point at which the molecules have stopped moving, which suggests that there must be a lowest possible temperature. We would not have come to the same conclusion if we had based our temperature measurement on the length of a metal rod since this will never be zero, no matter how cold it is.

Defining the scale

There is no problem having ice and water existing in equilibrium but to make the water boil at the same time, the pressure must be reduced.

All temperature scales are based on some physical property that changes with changing temperature. The absolute temperature scale is based on the pressure of a fixed mass of gas kept at constant volume. To define the size of any unit we need two fixed points (e.g. the metre could be defined in terms of the two ends of a metal rod).The absolute temperature scale is an exception to this rule, since the zero on the scale is absolute. In this case we only need one fixed point. A fixed point used to define a temperature scale is some observable event that always takes place at the same temperature; the commonly used examples are the melting and boiling of pure water at normal atmospheric pressure. The one used to define the absolute temperature scale is the triple point of water. This is the temperature at which water exists as solid, liquid and gas in equilibrium, 0.01 °C, almost the same as the freezing point.

Setting the size of the unit

To set the size of the unit we simply divide the difference between the fixed points into a convenient number of steps. The metre, for example, is divided into 100 centimetre steps. It might seem sensible to split the difference between the triple point and absolute zero into 100; however, this would make converting between Celsius and the new scale rather awkward. It is much better to choose the same division as used in the Celsius scale, then the units will be of equal size. If we look at Figure 3.26 we see that there are 273 °C between absolute zero and the triple point (about 0 °C) , so if we make the triple point 273 on the new scale, the two will be the same.

To clarify this, let us take an example.

Figure 3.27 shows the sort of apparatus that could be used to carry out this experiment. A sample of gas is cooled down to the triple point by placing it in a bath containing water, ice and steam in equilibrium. The pressure of the gas at this temperature is then measured to be 75kPa. To make the scale, this point is

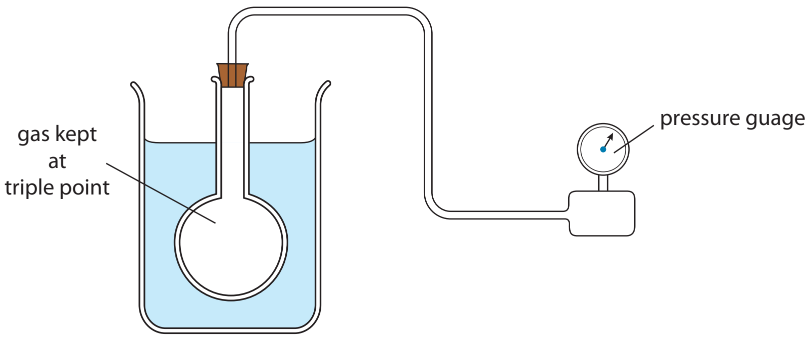
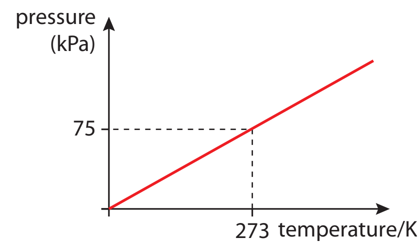


Figure 3.27

plotted on a graph as shown in Figure 3.28. By choosing the triple point to be 273 we make the gradient of this graph the same as Figure 3.26. This makes conversion from Celsius a simple matter of adding 273.

Figure 3.28 Graph used to define the Kelvin scale.



Converting between the Kelvin and Celsius scales

0 K = −273 °C

273 K = 0 °C

To convert from K to °C subtract 273.

To convert from °C to K add 273.

Since the size of the unit is the same, a *change* in temperature is the same in K and °C.

Example

Water is heated from 20 °C to 80 °C; this is a change of 80 − 20 = 60 °C

In kelvin, this is from 293 K to 353 K; this is a change of 353 − 293 = 60 K

For more precise calculations you should use a value of 273.15K = 0°C

Defining the state of a gas

Defining the state of a body means to give all the quantities necessary to enable someone else to recreate the exact conditions that you observe. A piece of metal can have many different temperatures and when its temperature is increased it expands. So to define the state of a given mass of metal you would need to quote its temperature and volume. To represent these states graphically a graph of temperature against volume could be plotted. All possible states would then lie on a straight line.

To define the state of a sample of gas you need to quote three quantities: pressure, volume and temperature. To represent these three quantities we would need to plot a three-dimensional graph. If we do this for an ideal gas we get a curved surface like the one shown in Figure 3.29. All possible states of the gas then lie on the curved surface.

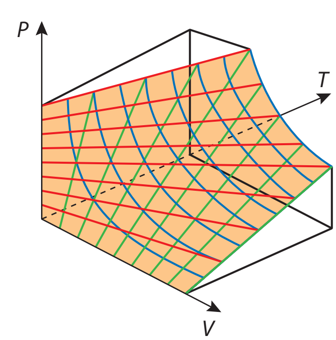


Figure 3.29

The equation of this surface is

$$PV = nRT$$

where n = the number of moles

R = the molar gas constant ($8.31\ \mathrm{J\,mol^{-1}\,K^{-1}}$)

This is called the equation of state for an ideal gas.

Worked example

The pressure of a gas inside a cylinder is 300 kPa. If the gas is compressed to half its original volume and the temperature rises from 23 °C to 323 °C, what will its new pressure be?

Solution

Using the ideal gas equation: $PV = nRT$

Rearranging: $\frac{PV}{T} = \text{consant}$

So $\frac{PV}{T}$ at the beginning $= \frac{PV}{T}$ at the end

$$\frac{PV}{T} \text{ at the beginning} = \frac{300\,000 \times V}{300}$$

$$\frac{PV}{T} \text{ at the end} = \frac{P \times V/2}{600}$$

Equating: $300\,000 \times \frac{V}{300} = \frac{P \times V/2}{600}$

$$P = 300\,000 \times 600 \times \frac{2}{300}$$

$$P = 1200\text{ kPa}$$

Exercises

15 The pressure of 10 m^3 of gas in a sealed container at 300 K is 250 kPa. If the temperature of the gas is changed to 350 K, what will the pressure be?

16 A container of volume 2 m^3 contains 5 moles of gas. If the temperature of the gas is 293 K.

(a) what is the pressure exerted by the gas?

(b) what is the new pressure if half of the gas leaks out?

17 A piston contains 250 cm^3 of gas at 300 K and a pressure of 150 kPa. The gas expands, causing the pressure to go down to 100 kPa and the temperature drops to 250 K. What is the new volume?

18 A sample of gas trapped in a piston is heated and compressed at the same time. This results in a doubling of temperature and a halving of the volume. If the initial pressure was 100 kPa, what will the final pressure be?

PV diagrams

It is rather difficult to draw the 3D graph that represents PV and T so to make life simpler we draw the view looking along the T axis. This is called a PV diagram and is shown in Figure 3.30.

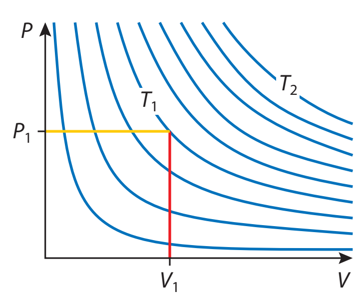


Figure 3.30 The curved surface of Figure.3.29 viewed in 2D.

You can imagine that the whole area of the PV diagram is covered in points. Each point represents the gas at a different state, in other words with different PV and T. If we join up all the points with equal pressure P_1 we get a horizontal line; if we join all points with equal volume V_1 we get a vertical line, and joining all points with equal temperature T_1 gives the blue curve shown.

The higher temperature curves are the ones further away from the origin, so $T_2 > T_1$. To explain this, consider a gas at volume V_1 pressure P_1 and temperature T_1. If the volume is kept constant as the temperature is increased, its pressure will increase. So its position on the diagram will move up the y-axis to a higher blue line. The blue lines are called isotherms, and although they are not always drawn on the PV diagram you must never forget that they are there.

Gas transformations

A gas can be heated, cooled, compressed and expanded. But, according to the equation of state for an ideal gas, whatever we do to the gas the value of $\frac{PV}{T}$ will remain the same. We can represent these changes on a PV diagram as illustrated in Figure 3.31.

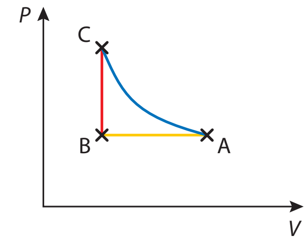


Figure 3.31

Constant pressure (isobaric)

The line A–B represents a constant pressure change. From A to B the volume is getting smaller, so this is a compression. When this happens, we can also deduce that the temperature must decrease, since the gas moves to a lower isotherm.

Constant volume (isochoric)

The line B–C represents a constant volume change. From B to C the pressure is increasing. This is because the temperature is increasing as can be deduced from the fact that the gas is changing to a higher temperature isotherm.

Constant temperature (isothermal)

The line C–A is an isotherm so represents a change at constant temperature. From C to A the volume is increasing, so this is an expansion.

Exercise

19 Figure 3.32 represents four different transitions performed on the same sample of a gas. For each transition (**a**, **b**, **c** and **d**) deduce whether each of P, V and T go up, down or stay the same.

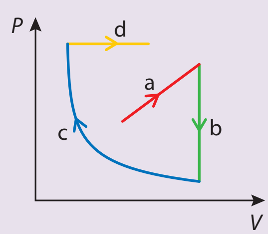


Figure 3.32

Real gases

If we compress a gas, the molecules get closer and closer together.

According to our simple kinetic theory there is no force between the molecules, so the only effect of compressing the gas is that the molecules become denser and therefore hit the walls more often, resulting in a proportional increase in pressure. However, as the gas molecules become very close to each other, the force between them is no longer negligible. As you try to push them closer, they push back. The pressure is no longer proportional to the volume but rises very steeply as the volume is reduced. What has happened is that the gas has changed to a liquid and that is what most real gases do, unless the temperature is very high. At high temperatures, no matter how much you compress the gas it will not turn to liquid.

Ideal gases and real gases
Ideal and real are not two types of gas. Ideal and real are the way the gas behaves. P, V and T for an ideal gas are related by the equation $PV = nRT$. Carbon dioxide behaves like an ideal gas at high temperatures and low pressures but not at low temperatures and high pressures.

3.5 Thermodynamic processes

Assessment statements

10.2.1 Deduce an expression for the work involved in a volume change of a gas at constant pressure.
10.2.2 State the first law of thermodynamics.
10.2.3 Identify the first law of thermodynamics as a statement of the principle of energy conservation.
10.2.4 Describe the isochoric (isovolumetric), isobaric, isothermal and adiabatic changes of state of an ideal gas.
10.2.5 Draw and annotate thermodynamic processes and cycles on P–V diagrams.
10.2.6 Calculate from a *P–V* diagram the work done in a thermodynamic cycle.
10.2.7 Solve problems involving state changes of a gas.

Thermodynamic system
Thermodynamics relates to a thermodynamic system – this is a collection of bodies that can do work on and exchange heat between each other. The car engine and the human body are both thermodynamic systems. In this course we will consider only the simple system of a gas trapped by a piston. However, the laws apply to all systems.

Energy and gas transformations

When dealing with the motion of a simple ball, we found that if we used the law of conservation of energy to solve problems, it was often simpler than going into the details of all the forces, acceleration etc. The same is true when dealing with the billions of particles that make up a gas – as long as the system is isolated, we can use the conservation of energy to predict the outcome of any transformation.
But before we can do that, we need to know in what way energy is involved when a gas changes state.

Internal energy

From the study of mechanics we know that a particle can possess two types of energy: PE and KE. According to our simple kinetic model of a gas there are no forces between the molecules. This means that to move a molecule around requires no work to be done (work done = force × distance) so moving the molecules will not result in a change in the PE. On the other hand, we know that the molecules are moving about in random motion – they therefore do possess KE. The sum of all the KE of all the molecules is called the *internal energy*.

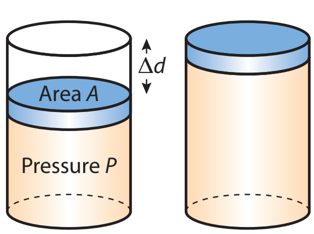


Figure 3.33 A gas expands at constant pressure.

Work done

Work is done when the point of application of a force moves in the direction of the force. If the pressure of a gas pushes a piston out, then the force exerted on the piston is moving in the direction of the force, so work is done. The example in Figure 3.33 is of a gas expanding at constant pressure. In this case, the force exerted on the piston $= P \times A$. The work done when the piston moved distance Δd is therefore given by:

$$\text{Work done} = P \times A \times \Delta d$$

but $A\Delta d$ is the change in volume ΔV, so

$$\text{Work done} = P\Delta V$$

Figure 3.34 is the PV graph for this constant pressure expansion. From this we can see that the work done is given by the area under the graph. This is true for all processes.

> **Sign of work**
> When a gas does work, it is pushing the piston out; this is positive.
> If work is done on the gas then something must be pushing the piston in. This is negative work.

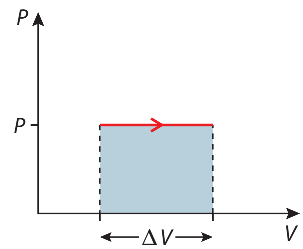


Figure 3.34

Heat

Heat is the name given to the energy that is exchanged when a hot body is in contact with a cold one. Adding heat to a gas can increase its internal energy and allow it to do work.

The first law of thermodynamics

According to the law of conservation of energy, energy can neither be created nor destroyed, so the amount of heat, Q, added to a gas must equal the work done by the gas, W, plus the increase in internal energy, ΔU. This is so fundamental to the way physical systems behave that it is called the *first law of thermodynamics*. This can be written in the following way

$$Q = \Delta U + W$$

> **First law (simple version)**
> If a gas expands and gets hot, heat must have been added.

This would be nice and easy if the only thing a gas could do is gain heat, get hot and do work. However, heat can be added and lost, work can be done by the gas and on the gas and the internal energy can increase and decrease. To help us understand all the different possibilities, we will use the PV diagram to represent the states of a gas.

Using *PV* diagrams in thermodynamics

We have seen how a PV diagram enables us to see the changes in P, V and T that take place when a gas changes from one state to another. It also tells us what energy changes are taking place. If we consider the transformation represented in Figure 3.35 we can deduce that when the gas changes from A to B:

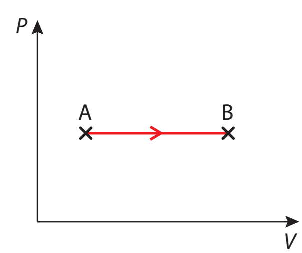


Figure 3.35

1 Since the volume is increasing, the gas is doing work (W is positive).

2 Since the temperature is increasing, the internal energy is increasing (ΔU is positive).

If we then apply the first law $Q = \Delta U + W$ we can conclude that if both ΔU and W are positive then Q must also be positive, so heat must have been added.

This is a typical example of how we use the PV diagram with the first law; we use the diagram to find out how the temperature changes and whether work is done *by* the gas or *on* the gas, and then use the first law to deduce whether heat is added or lost.

> **Examiner's hint:**
> Change in **volume** tells us whether work is done by the gas or on it.
> Change in **temperature** tells us whether the internal energy goes up or down.
> Change in **pressure** is not interesting.

Examples

Constant pressure contraction (isobaric)

The example of Figure 3.35 was an isobaric expansion; if we reverse this transformation (B $\rightarrow$ A) we get an isobaric compression. In this case:

1 Temperature decrease implies that the internal energy decreases (ΔU = negative).

2 Volume decrease implies that work is done on the gas (W = negative).

Applying the first law, $Q = \Delta U + W$, tells us that Q is also negative, so heat is lost.

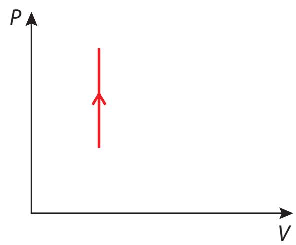


Figure 3.36 An isochoric transformation.

Constant volume increase in temperature (isochoric)

Figure 3.36 is the PV graph for a gas undergoing a constant volume transformation. From the graph we can deduce that:

1 The volume isn't changing, so no work is done ($W = 0$).

2 The gas changes to a higher isotherm so the temperature is increasing; this means that the internal energy is increasing (ΔU = positive).

Applying the first law $Q = \Delta U + W$ we can conclude that $Q = \Delta U$ so if ΔU is positive then Q is also positive – heat has been added.

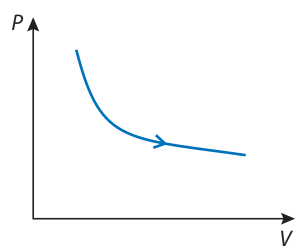


Figure 3.37 A isothermal expansion.

Isothermal expansion

When a gas expands isothermally the transformation follows an isothermal as shown in Figure 3.37. From this PV diagram we can deduce that:

1 The temperature doesn't change so there is no change in internal energy ($\Delta U = 0$)

2 The volume increases so work is done by the gas (W = positive).

Applying the first law, $Q = \Delta U + W$, we conclude that $Q = W$ so heat must have been added. The heat added enables the gas to do work.

Adiabatic contraction

An adiabatic transformation is one where no heat is exchanged ($Q = 0$). This transformation is represented by the red line in Figure 3.38. It is much steeper than an isothermal. From the PV diagram we can deduce that:

1 The volume is reduced so work is done on the gas (W = negative).

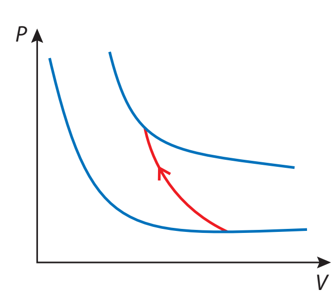
2 The temperature increases so the internal energy increases (ΔU = positive).

We also know that $Q = 0$ so if we apply the first law $Q = \Delta U + W$, we get

$$0 = \Delta U - W$$

Rearranging gives $\Delta U = W$, so the work done on the gas goes to increase the internal energy.

Figure 3.38 Adiabatic contraction.



Exercises

20 Calculate the work done by the gas when it expands from a volume of 250 cm^3 to 350 cm^3 at a constant pressure of 200 kPa.

21 Estimate the work done by the gas that undergoes the transformation shown in Figure 3.39.

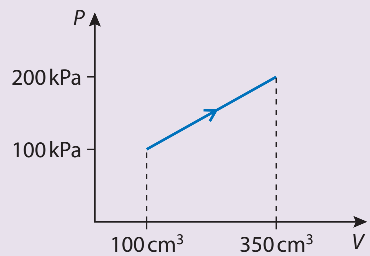


Figure 3.39

● Examiner's hint:
You can't easily tell if a curve on a *PV* graph is adiabatic or isothermal - you have to be told by the examiner. If you have both on a diagram then the adiabatic is the steeper one.

Thermodynamic cycles

A thermodynamic cycle is represented by a closed loop on a *PV* diagram as in Figure 3.40. In this example the cycle is clockwise so the sequence of transformations is:

A–B isochoric temperature rise

B–C isobaric expansion

C–D isochoric temperature drop

D–A isobaric compression.

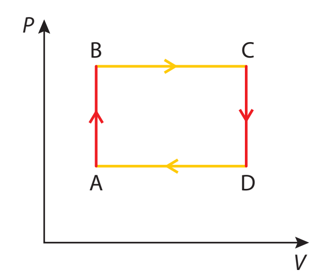


Figure 3.40 A thermodynamic cycle.

In the process of completing this cycle, work is done on the gas from D to A and the gas does work from B to C. It is clear from the diagram that the work done by the gas is greater than the work done on the gas (since the area under the graph is greater from B to C than from D to A) so net work is done. What we have here is an engine; heat is added and work is done. Let us look at this cycle more closely.

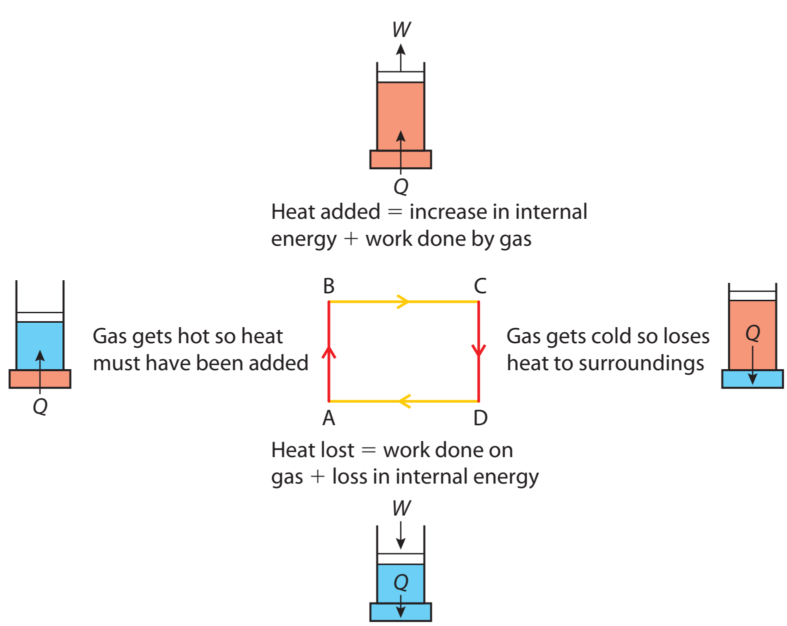


Figure 3.41 An example of a thermodynamic cycle, the red and blue rectangles placed under the piston represent hot and cold bodies used to add and take away heat.

Net work done
The net work done during a cycle is the difference between the work done by the gas and the work done on the gas. This is equal to the area enclosed by the cycle on the *PV* diagram.

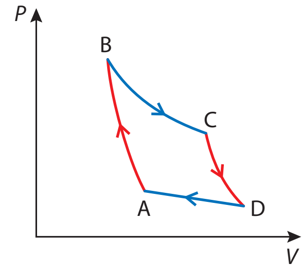


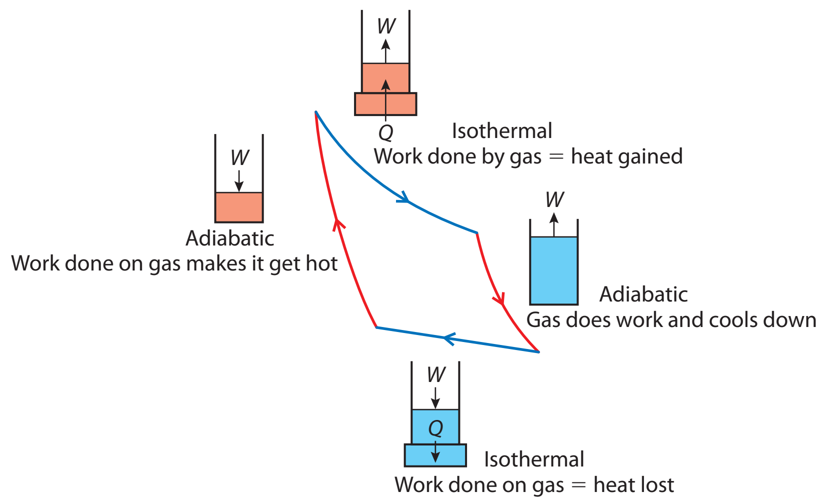
Figure 3.42 The Carnot cycle.

The secret to the operation of all heat engines is that the gas is cooled down before it is compressed back to its original volume. The cold gas is easier to compress than a hot one so when the gas is hot it does work, but it's reset when it's cold.

The Carnot cycle

The Carnot cycle is a cycle of only isothermal and adiabatic changes as shown in Figure 3.42. It might look more complicated than the previous example but in terms of thermodynamics, it is simpler. In this cycle, work is done by the gas from B to D and work is done on the gas from D to B. Since the work done by the gas is greater, this cycle is operating as an engine. The details of this cycle are shown in Figure 3.43, but the principle is the same as all engines; the piston pushes out when the gas is hot and is pushed back in when it's cold.

Figure 3.43 The Carnot cycle in detail.

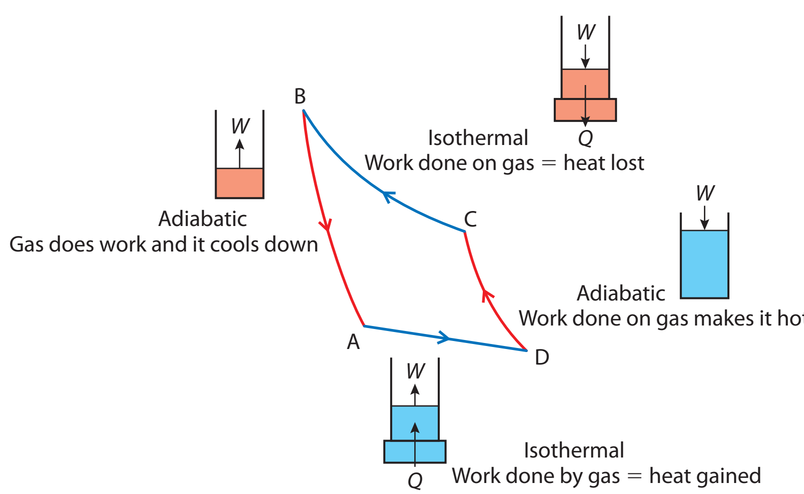


It is much easier to visualise what is happening if you look at a simulation. To do this, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 3.3.

The reverse cycle

Let us consider what would happen if the Carnot cycle was operated in reverse. The details of this are shown in Figure 3.44

Figure 3.44 The reverse Carnot cycle.



The heat pump

A heat pump is used to extract heat from the cold air outside and give it to the inside of a house. It works in exactly the same way as a refrigerator.

The interesting thing about this cycle is that heat is lost to the hot body during the isothermal compression (C to B) and gained from the cold body during the isothermal expansion (A to D). So heat has been taken from something cold and given to something hot. This is what a refrigerator does - it takes heat from the cold food inside and gives it to the warm room. To make this possible, work must be done on the gas (D to C) so that it gets hot enough to give heat to the hot body.

Exercises

22 250 cm^3 of gas at 300 K exerts a pressure of 100 kPa on its container; call this state A. It undergoes the following cycle of transformations:

(i) an isobaric expansion to 500 cm^3 (state B)

(ii) an isochoric transformation to a pressure of 200 kPa (state C)

(iii) an isobaric contraction back to 250 cm^3 (state D)

(iv) an isochoric transformation back to state A.

(a) Sketch a *PV* diagram representing this cycle, labelling the states A, B , C and D.

(b) Use the ideal gas equation to calculate the temperature at B, C and D.

(c) Calculate the amount of work done by the gas.

(d) Calculate the amount of work done on the gas

(e) What is the net work done during one cycle?

23 Figure. 3.45 represents a Carnot cycle.
The areas of the coloured regions are as follows

A – 50 J
B – 45 J
C – 40 J
D – 35 J
E – 150 J

If the cycle is performed clockwise, how much work is done:

(a) during the isothermal expansion?

(b) during the adiabatic compression?

(c) by the gas?

(d) on the gas?

(e) in total?

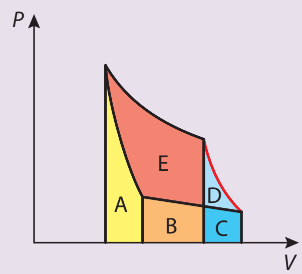


Figure 3.45

3.6 The second law of thermodynamics

Assessment statements

10.3.1 State that the second law of thermodynamics implies that thermal energy cannot spontaneously transfer from a region of low temperature to a region of high temperature.

10.3.2 State that entropy is a system property that expresses the degree of disorder in the system.

10.3.3 State the second law of thermodynamics in terms of entropy changes.

10.3.4 Discuss examples of natural processes in terms of entropy changes.

Isothermal expansions

When looking at examples of the first law of thermodynamics, we used isothermal transformations as a simple example. However, if we look at these processes closely we find that they are not possible. Consider a gas enclosed by a piston shown in Figure 3.46 To enable this to expand we could put the gas in contact with a hot source. Heat would then flow into the gas increasing the KE of the molecules. The increased pressure would then push the cylinder out doing work against the surroundings, and the work done would cause the temperature of the gas to drop. For this expansion to be isothermal the temperature would have to stay constant.

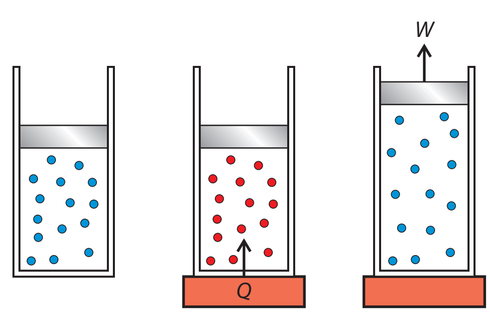
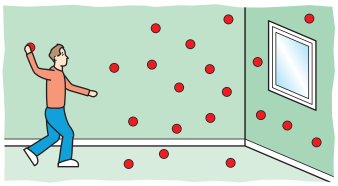
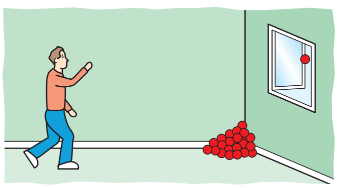


Figure 3.46





The energy you give to the gas must be directly transferred to the piston without increasing the average KE of the gas. The problem is that when the energy is given to the gas molecules, the energy spreads out; this is because all the gas molecules collide into one another. Once the energy is spread out it is impossible to get it all back again. Imagine you are in a room full of perfectly elastic red balls that are bouncing around the room in random motion. You are standing on one side of the room and on the other side is a window. You want to open the window by throwing one of the balls at it. The problem is that every time you throw the ball it hits all the others, and the KE you give it is shared amongst all the other balls and it never reaches its destination.

According to the first law, as long as energy is conserved, anything is possible; but now we see that some things aren't possible. So, to take into account the fact that some things aren't possible, we add the second law:

It is not possible to convert heat completely into work.

Absolute zero

There is one solution to this problem and you can work it out by considering the situation in the room with the rubber balls. If you catch all the rubber balls and put them on the floor you could then take one of the balls and throw it at the window. It would now be able to fly across the room uninterrupted and push the window open. In terms of the gas, this is equivalent to reducing the temperature to zero kelvin, so according to our model, an isothermal expansion is only possible at 0 K. The problem of doing this in the room is that every time you catch a ball and put it down it would get hit by one of the others. You would never be able to stop all the balls, and for the same reason it is not possible to reach 0 K.

Implications of the second law

The second law does not only tell us that isothermal processes cannot take place but it tells us something fundamental about how the physical world behaves. Here are some examples.

Heat flows from hot to cold

If you have some hot gas next to a cold gas, the heat will always travel from hot to cold, never the other way round. We can explain this by using the kinetic model; the gas is made up of randomly moving particles that collide with each other, and the hot gas has faster moving particles. When a fast particle hits a slow one, energy is transferred from the fast one to the slow one, not the other way around.

All the air in a room never goes out of the window

If you are sitting in a room with the window open, is it possible that suddenly all the air molecules could fly out of the window leaving you in a room with no air? Again, if we consider the kinetic model, all the particles are moving in random motion, so for all the molecules to move out of the window, something would have to push them in the same direction. Without that external force this cannot happen.

Spreading out of energy

An alternative way of quoting the second law is to say that energy always spreads out. This can be illustrated by considering molecules injected into a container with identical velocity, as shown in Figure 3.47.

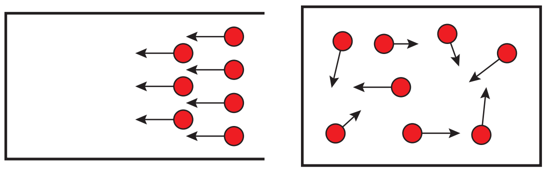


Figure 3.47 Molecules start ordered but end up in random motion.

After they have hit the walls they start to collide with each other. Once this happens, the energy of the individual molecules changes; some will gain energy, some will lose energy, and their motion changes from ordered to disordered. They will never have the same energy again. This is the way of the universe; energy always spreads out.

To view the simulation 'gas properties', visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 3.4.

Entropy

The second law of thermodynamics is about the spreading out of energy. This can be quantified by using the quantity *entropy*.

The change of entropy is ΔS, when a quantity of heat flow into a body at temperature T is equal to $\frac{Q}{T}$.

$$\Delta S = \frac{Q}{T}$$

The unit of entropy is J K^{-1}.

For example, consider the situation of a 1 kg block of ice melting in a room that is at a constant temperature 300 K. To melt the block of ice, it must gain 3.35×10^5 J of energy. Ice melts at a constant 273 K so:

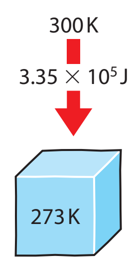


Figure 3.48

The gain in entropy of the ice $= \frac{3.35 \times 10^5}{273} = 1.23 \times 10^3 \text{ J K}^{-1}$

The loss of entropy by the room $= \frac{3.35 \times 10^5}{300} = 1.12 \times 10^3 \text{ J K}^{-1}$

We can see from this that the entropy has increased.

Entropy always increases in any transfer of heat since heat always flows from hot bodies to cold bodies. We can therefore rewrite the second law in terms of entropy.

In any thermodynamic process the total entropy always increases.

Entropy is a measure of how spread out or disordered the energy has become. Saying entropy has increased implies that the energy has become more spread out.

Examples

Even though locally entropy can decrease, the total entropy of a system must always increase. Sometimes it's not so easy to see this, but remember it must always happen.

The second law is a way of describing the way the bodies in the universe behave. It is often said that something can't happen because the second law says so, but this isn't really true. Some things don't happen because the physical world simply isn't like that, not because the law doesn't allow it.

1 A falling stone
When a stone is held above the ground, all the molecules in the stone are lifted to approximately the same height and therefore have the same PE; this energy is very ordered. If the stone is dropped and hits the ground, this PE is converted into thermal energy in the stone and ground. Each molecule will gain some PE and some KE in a fairly random fashion. The energy is now disordered, and entropy has increased. As time goes on, entropy continues to increase, as this thermal energy spreads out in the ground. It is not possible to collect all this energy back into the stone and have it thrown back into the air; this would be against the second law.

2 The refrigerator
The food in a refrigerator gets cold. As things get cold, the molecules become more ordered, and the entropy in the fridge therefore decreases. According to the second law, entropy must always increase, so the room must gain heat.

3 The petrol engine
In a petrol engine, petrol is burnt to produce heat, which causes a gas to expand, enabling it to do work. If the engine is used to lift a load then the disordered energy in the hot gas has been converted to ordered energy in the lifted load. According to the second law, entropy must increase, so there must be some disorder created. When any engine is used, some heat is lost to the surroundings, and this is where the disorder is created. From this we can deduce that it is impossible to make an engine that does not give heat to something cold; this means an engine can never be 100% efficient.

Exercise

24 500 J of heat flows from a hot body at 400 K to a colder one at 250 K.
(a) Calculate the entropy change in
(i) the hot body
(ii) the cold body.
(b) What is the total change in entropy?

25 Use the second law to explain why heat is released when an electric motor is used to lift a heavy load.

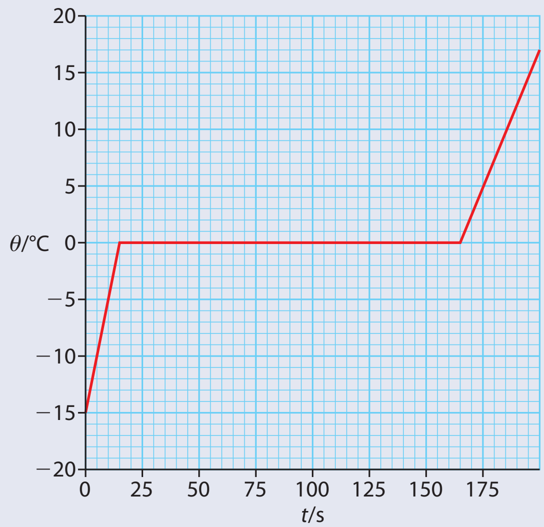
Practice questions

1 This question is about the change of phase (state) of ice.

A quantity of crushed ice is removed from a freezer and placed in a calorimeter. Thermal energy is supplied to the ice at a constant rate. To ensure that all the ice is at the same temperature, it is continually stirred. The temperature of the contents of the calorimeter is recorded every 15 seconds.

The graph at the top of the next page shows the variation with time t of the temperature θ of the contents of the calorimeter. (Uncertainties in the measured quantities are not shown.)

(a) On the graph, mark with an X, the data point on the graph at which all the ice has just melted. (1)
(b) Explain, with reference to the energy of the molecules, the constant temperature region of the graph. (3)



The mass of the ice is 0.25 kg and the specific heat capacity of water is 4200 J kg^{-1} K^{-1}.

(c) Use these data and data from the graph to

(i) deduce that energy is supplied to the ice at the rate of about 530 W (3)

(ii) determine the specific heat capacity of ice (3)

(iii) determine the specific latent heat of fusion of ice. (2)

2 This question is about thermal physics.

(a) Explain why, when a liquid evaporates, the liquid cools unless thermal energy is supplied to it. (3)

(b) State two factors that cause an increase in the rate of evaporation of a liquid. (2)

(c) A mass of 350 g of water at a temperature of 25 °C is placed in a refrigerator that extracts thermal energy from the water at a rate of 86 W.
Calculate the time taken for the water to become ice at −5.0 °C. (3)

Specific heat capacity of ice = 2.1×10^3 J kg^{-1} K^{-1}
Specific heat capacity of water = 4.2×10^3 J kg^{-1} K^{-1}
Specific latent heat of fusion of ice = 3.3×10^5 J kg^{-1}

3 Explain, in terms of the behaviour of the molecules of an ideal gas, why the pressure of the gas rises when it is heated at constant volume. (3)

4 This question is about the phase (state) changes of the element lead.
A sample of lead has a mass of 0.50 kg and a temperature of 27 °C. Energy is supplied to the lead at the rate of 1.5 kW. After 0.2 minutes of heating it reaches its melting point temperature of 327 °C. After heating for a further 3 minutes, all the lead has become liquid.

(a) Assuming that all the energy goes into heating the lead, calculate a value for the

(i) specific heat capacity of lead (3)

(ii) latent heat of fusion of lead. (2)

(b) Energy continues to be supplied to the lead. Sketch a graph to show how the temperature of the lead varies with time from the start of heating to some 5 minutes after the time when all the lead has become liquid. Indicate on the graph the time at which it starts to melt and the time when it has become liquid. (2)
(You are not expected to have accurate scales; this is just a sketch graph.)

5 This question is about modelling the thermal processes involved when a person is running. When running, a person generates thermal energy but maintains approximately constant temperature.

(a) Explain what *thermal energy* and *temperature* mean. Distinguish between the two concepts. (4)

The following simple model may be used to estimate the rise in temperature of a runner, assuming no thermal energy is lost.

A closed container holds 70 kg of water, representing the mass of the runner. The water is heated at a rate of 1200 W for 30 minutes. This represents the energy generation in the runner.

(b) **(i)** Show that the thermal energy generated by the heater is 2.2×10^6 J. (2)

(ii) Calculate the temperature rise of the water, assuming no energy losses from the water. (3)

The specific heat capacity of water is 4200 J kg^{-1} K^{-1}.

(c) The temperature rise calculated in **(b)** would be dangerous for the runner. Outline three mechanisms, other than evaporation, by which the container in the model would transfer energy to its surroundings. (6)

(d) A further process by which energy is lost from the runner is the evaporation of sweat.

(i) Describe, in terms of molecular behaviour, why evaporation causes cooling. (3)

(ii) Percentage of generated energy lost by sweating: 50%

Specific latent heat of vaporization of sweat: 2.26×10^6 J kg^{-1}

Using the information above, and your answer to **(b) (i)**, estimate the mass of sweat evaporated from the runner. (3)

(iii) State and explain two factors that affect the rate of evaporation of sweat from the skin of the runner. (4)

6 A gas is contained in a cylinder fitted with a piston as shown below.

When the gas is compressed rapidly by the piston, its temperature rises because the molecules of the gas

A are squeezed closer together.
B collide with each other more frequently.
C collide with the walls of the container more frequently.
D gain energy from the moving piston. (1)

7 The Kelvin temperature of an ideal gas is a measure of the

A average speed of the molecules.
B average momentum of the molecules.
C average kinetic energy of the molecules.
D average potential energy of the molecules. (1)

8 The temperature of an ideal gas is reduced. Which one of the following statements is true?

A The molecules collide with the walls of the container less frequently.
B The molecules collide with each other more frequently.
C The time of contact between the molecules and the wall is reduced.
D The time of contact between molecules is increased. (1)

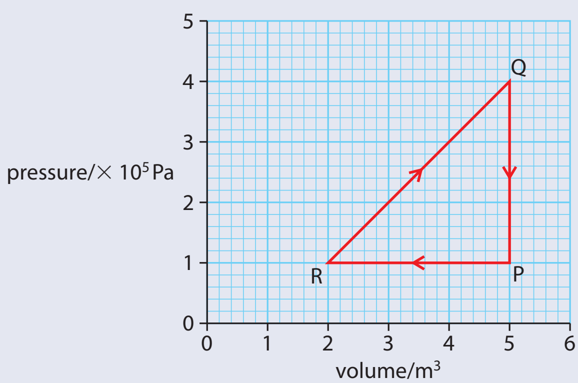
9 When a gas in a cylinder is compressed at constant temperature by a piston, the pressure of the gas increases. Consider the following three statements.

I The rate at which the molecules collide with the piston increases.
II The average speed of the molecules increases.
III The molecules collide with each other more often.

Which statement(s) correctly explain the increase in pressure?

A I only
B II only
C I and II only
D I and III only (1)

10 The graph below shows the variation with volume of the pressure of a system.

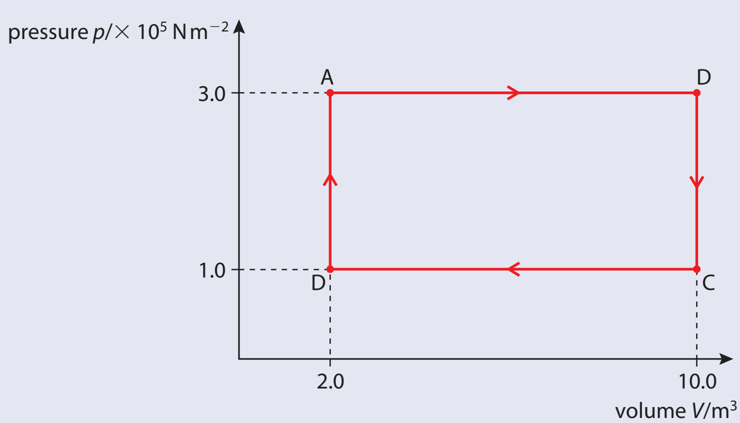


The work done in compressing the gas from R to P is

A 5.0×10^5 J.
B 4.5×10^5 J.
C 3.0×10^5 J.
D 0. (1)

11 This question is about the thermodynamics of a heat engine.

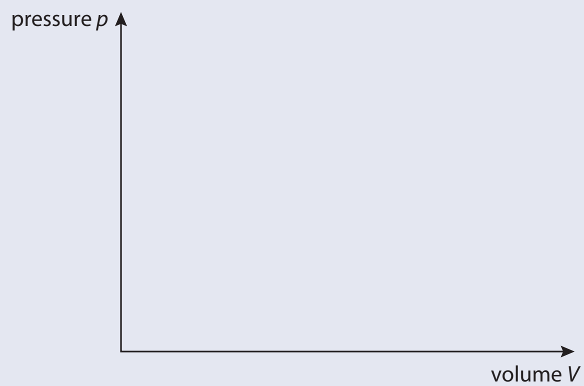
In an idealized heat engine, a fixed mass of a gas undergoes various changes of temperature, pressure and volume. The *p–V* cycle (A→B→C→D→A) for these changes is shown in the diagram below.



(a) Use the information from the graph to calculate the work done during **one** cycle. (2)

(b) During one cycle, a total of 1.8×10^6 J of thermal energy is ejected into a cold reservoir. Calculate the efficiency of this engine. (2)

(c) Using the axes below, sketch the p–V changes that take place in the fixed mass of an ideal gas during one cycle of a Carnot engine. *(Note this is a sketch graph – you do not need to add any values.)*



(2)

(d) **(i)** State the names of the **two** types of change that take place during one cycle of a Carnot engine. (2)

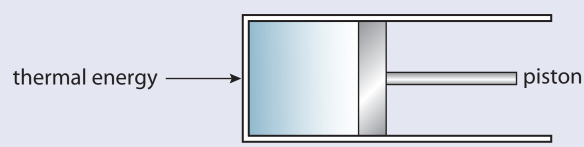
(ii) Add labels to the above graph to indicate which parts of the cycle refer to which particular type of change. (2)

(Total 10 marks)

12 This question is about thermodynamic processes.

(a) Distinguish between an *isothermal* process and an *adiabatic* process as applied to an ideal gas. (2)

An ideal gas is held in a container by a moveable piston and thermal energy is supplied to the gas such that it expands at a constant pressure of 1.2×10^5 Pa.



The initial volume of the container is 0.050 m^3 and after expansion the volume is 0.10 m^3. The total energy supplied to the gas during the process is 8.0×10^3 J.

(b) **(i)** State whether this process is **either** isothermal **or** adiabatic **or** neither. (1)

(ii) Determine the work done by the gas. (1)

(iii) Hence calculate the change in internal energy of the gas. (2)

(Total 6 marks)

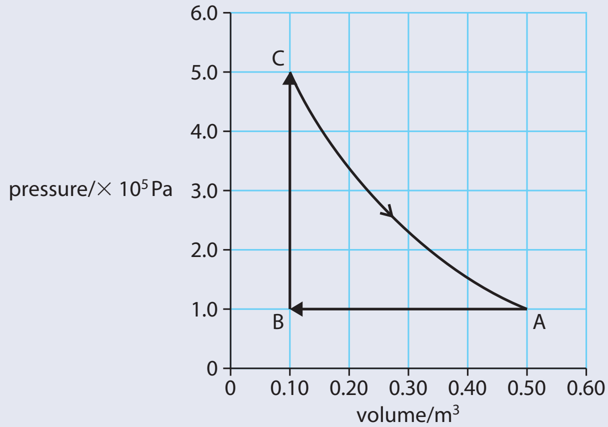
13 This question is about a heat engine.

A certain heat engine uses a fixed mass of an ideal gas as a working substance. The graph opposite shows the changes in pressure and volume of the gas during one cycle ABCA of operation of the engine.

(a) For the part A $\longrightarrow$ B of the cycle, explain whether

(i) work is done **by** the gas or work is done **on** the gas. (1)

(ii) thermal energy (heat) is absorbed **by** the gas or is ejected **from** the gas to the surrounding. (1)



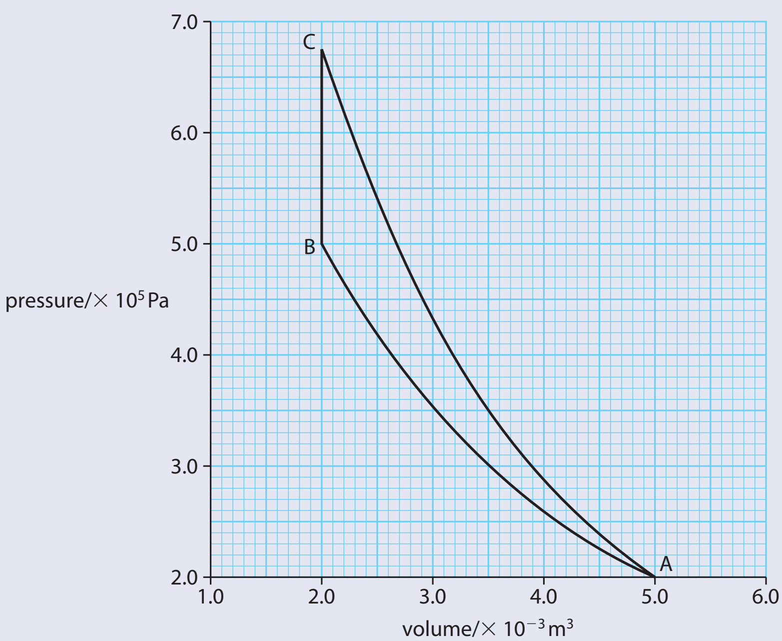
(b) Calculate the work done during the change A → B. (2)

(c) Use the graph to estimate the total work done during one cycle. (2)

(d) The total thermal energy supplied to the gas during one cycle is 120 kJ. Estimate the efficiency of this heat engine. (2)

(Total 8 marks)

14 This question is about *p–V* diagrams. The graph below shows the variation with volume of the pressure of a fixed mass of gas when it is compressed adiabatically and also when the same sample of gas is compressed isothermally.



(a) State and explain which line, AB or AC, represents the isothermal compression. (2)

(b) On the graph, shade the area that represents the difference in work done in the adiabatic change and in the isothermal change. (1)

(c) Determine the difference in work done, as identified in **(b)**. (3)

(d) Use the first law of thermodynamics to explain the change in temperature during the adiabatic compression. (3)

(Total 9 marks)

4 Simple harmonic motion and waves

4.1 Kinematics of simple harmonic motion

Assessment statements

4.1.1 Describe examples of oscillations.

4.1.2 Define the terms *displacement*, *amplitude*, *frequency*, *period* and *phase difference*.

4.1.3 Define *simple harmonic motion* (SHM) and state the defining equation as $a = -\omega^2 x$.

4.1.4 Solve problems using the defining equation for SHM.

4.1.5 Apply the equations $v = v_0 \sin \omega t$, $v = v_0 \cos \omega t$, $v = \pm\omega\sqrt{(x_0{}^2 - x^2)}$, $x = x_0 \cos \omega t$ and $x = x_0 \sin \omega t$ as solutions to the defining equation for SHM.

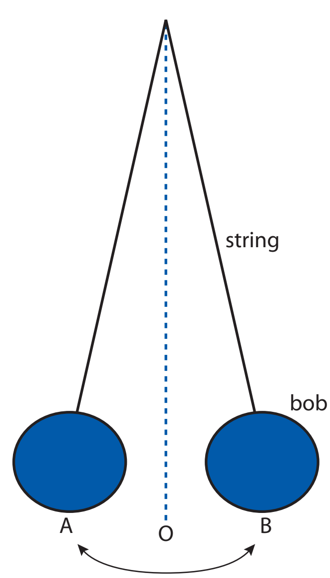
4.1.6 Solve problems, both graphically and by calculation, for acceleration, velocity and displacement during SHM.

A swing is an example of oscillatory motion.

Oscillations

In this section we will derive a mathematical model for an oscillating or vibrating body. There are many different examples of naturally occurring oscillations but they don't all have the same type of motion. We are going to consider the simplest form of oscillation: simple harmonic motion. The most common example of this is a pendulum (Figure 4.1). Before we start to model this motion, we need to define some new terms and quantities.

Figure 4.1 The simple pendulum swings from A to B and back.



Cycle

One cycle is defined as one complete oscillation of the pendulum (A–B–A). The term *cycle* is also used to describe circular motion; one cycle is one complete circle or 2π radians.

Equilibrium position

The equilibrium position is the position where the pendulum bob would rest if not disturbed – this is position O.

Amplitude (x_0)

The amplitude is defined as the maximum displacement from the equilibrium position, this is distance OB or OA.

Unit: metre

Measuring time period

To reduce the uncertainties when measuring the time period of an oscillation, it is best to measure the time for many oscillations e.g. 10. The time period is found by dividing this time by the number of oscillations.

Time period (T)

The time period is the time taken for one complete cycle.

Unit: second

Frequency (f)

The frequency is the number of cycles that the pendulum makes per unit time. This is equal to 1/time period.

Unit: s^{-1} or hertz (Hz)

Angular frequency (ω)

The angular frequency is found by multiplying f by 2π ($\omega = 2\pi f$). This quantity is normally used when describing circular motion. An angular frequency of 2π rads s^{-1} means that a body makes one revolution per second. However, it is also used to describe an oscillation, 2π being equivalent to one complete cycle.

Unit: s^{-1} or hertz (Hz)

Worked example

A pendulum completes 10 swings in 8 s. What is the angular frequency?

Solution

There are 10 swings in 8 seconds, so each swing takes 0.8 s.

Time period $= 0.8$ s.

Frequency $= \frac{1}{T} = \frac{1}{0.8} = 1.25$ Hz

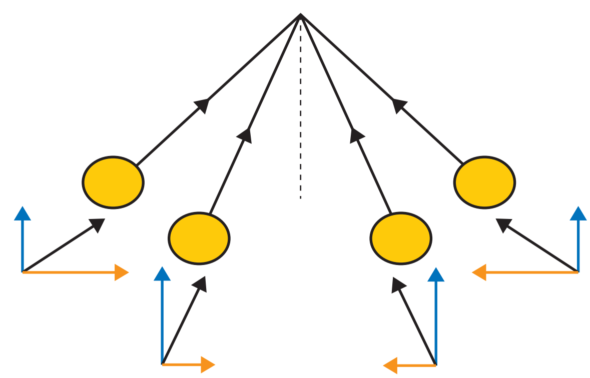
Angular frequency $\omega = 2\pi f = 2\pi \times 1.25 = \mathbf{7.8\ rad\ s^{-1}}$

Analysing oscillations

To make a model of oscillatory motion, we will analyse two different oscillations and see if there are any similarities.

The pendulum

Figure 4.2 As the angle increases, the horizontal component of tension increases, but is always pointing towards the centre.

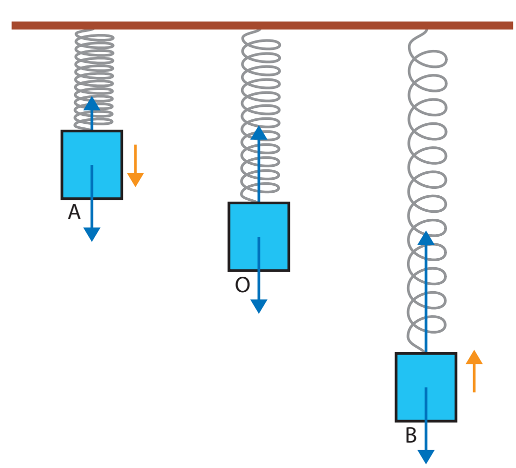


When a pendulum bob is pushed to one side and released, it will swing back down. The reason for this can be understood by drawing the forces acting on the bob. In Figure 4.2, we can see that when the string makes an angle to the vertical, the tension has a component in the horizontal direction; this component causes the bob to accelerate back towards the middle. As the bob swings down, the angle of the string gets smaller, and the horizontal component decreases. The horizontal acceleration of the bob is proportional to the horizontal force, so we can therefore deduce that the horizontal acceleration is proportional to the displacement from the lowest point.

When the ball reaches the lowest position, it is travelling at its maximum speed. It passes through this position and continues to swing up on the other side. The horizontal component of the tension is now acting in the other direction. This is in the *opposite* direction to the motion so will slow the bob down. We can conclude that no matter where the bob is, its acceleration is always directed towards O.

Mass on a spring

Figure 4.3 The tension increases as the spring is stretched. The resultant (red) increases with increased distance from the centre and is always directed towards the centre.



If a mass hanging on the end of a spring is lifted up and released, it will bounce up and down as in Figure 4.3. The forces acting on the mass are *weight* and the *tension in the spring*. The weight is always the same but the tension depends on how far the spring is stretched. (When you stretch a spring, the tension is proportional to the extension).

At A, the spring is short, so the tension will be small; the weight will therefore be bigger than the tension, so the resultant force will be downwards.

As the mass passes through the middle point, the forces will be balanced.

At B, the spring is stretched, so the tension is large; the tension will therefore be greater than the weight, so the resultant force will be upwards.

Again we can see that the acceleration is proportional to the displacement from the central point and always directed towards it.

This type of motion is called *simple harmonic motion* or SHM.

SHM
The acceleration is proportional to the distance from a fixed point.
The acceleration is always directed towards a fixed point.

Exercises

1 State whether the following are examples of simple harmonic motion.

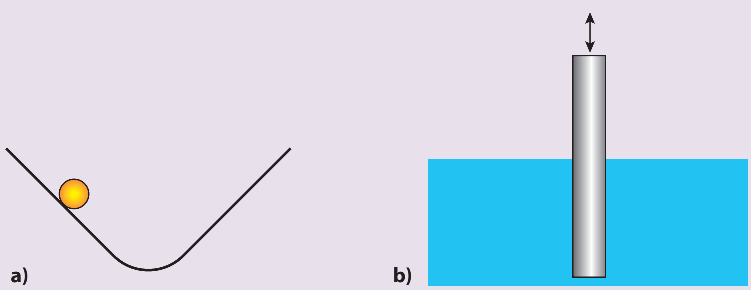


Figure 4.4

(a) A ball rolling up and down on a track (Figure 4.4a).
(b) A cylindrical tube floating in water when pushed down and released (Figure 4.4b).
(c) A tennis ball bouncing back and forth across the net.
(d) A bouncing ball.

2 A pendulum completes 20 swings in 12 s. What is
(a) the frequency?
(b) the angular frequency?

Graphical treatment

To analyse the oscillation further, we can plot graphs for the motion. In this example, we will consider a mass on a spring, but we could choose any simple harmonic motion.

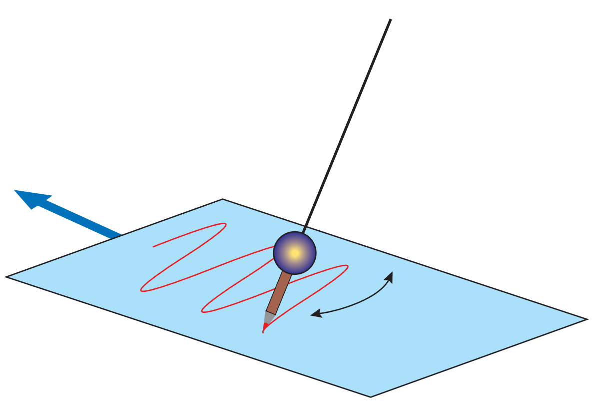


Figure 4.5 You can plot a displacement–time graph by attaching a pen to a pendulum and moving paper beneath it at a constant velocity.

To see how the equation fits the graph we can put some numbers into the equation.
In this example, the time period = 4 s
Therefore $f = \frac{1}{4} = 0.25$ Hz
Angular frequency $= 2\pi f = 0.5\pi$
So displacement $= 2\cos(0.5\pi t)$
Calculating displacement at different times gives:
$t = 1$ s $y = 2\cos(\pi/2) = 0$ cm
$t = 2$ s $y = 2\cos(\pi) = -2$ cm
$t = 3$ s $y = 2\cos(3\pi/2) = 0$ cm
$t = 4$ s $y = 2\cos(2\pi) = 2$ cm

Displacement–time

As before, O is the equilibrium position and we will take this to be our position of zero displacement. Above this is positive displacement and below is negative.

At A, the mass has maximum positive displacement from O.

At O, the mass has zero displacement from O.

At B, the mass has maximum negative displacement from O.

We can see that the shape of this displacement–time graph is a cosine curve.

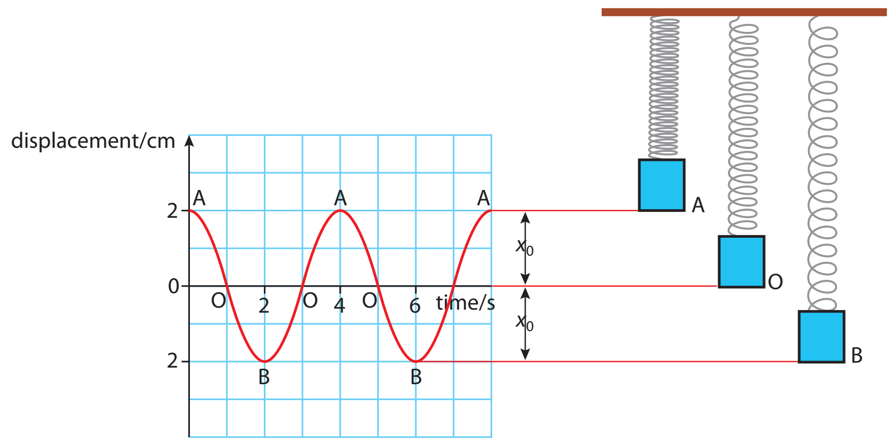


Figure 4.6 Displacement–time graph.

The equation of this line is $x = x_0 \cos \omega t$,

where x_0 *is* the maximum displacement and ω is the angular frequency.

Velocity–time

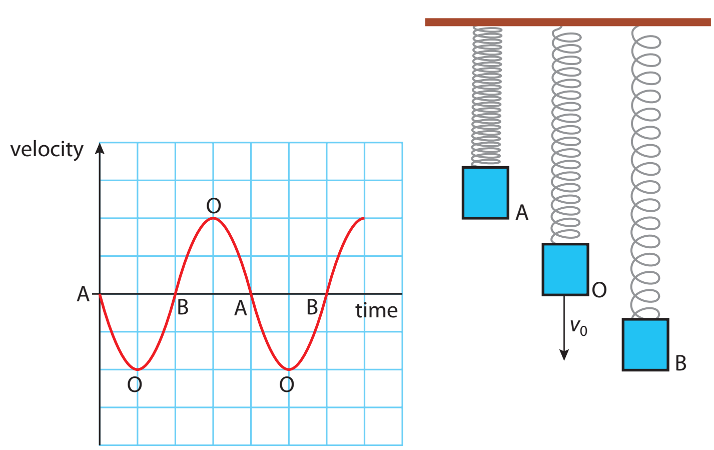
From the gradient of the displacement–time graph (Figure 4.6), we can calculate the velocity.

At A, gradient = 0 so velocity is zero.

At O, gradient is negative and maximum, so velocity is down and maximum.

At B, gradient = 0 so velocity is zero.

Figure 4.7 Velocity–time graph.



The equation of this line is $v = -v_0 \sin \omega t$ where v_0 is the maximum velocity.

Acceleration–time

From the gradient of the velocity–time graph (Figure 4.7) we can calculate the acceleration.

At A, the gradient is maximum and negative so acceleration is maximum and downwards.

At O, the gradient is zero so acceleration is zero.

At B, the gradient is maximum and positive so the acceleration is maximum and upwards.

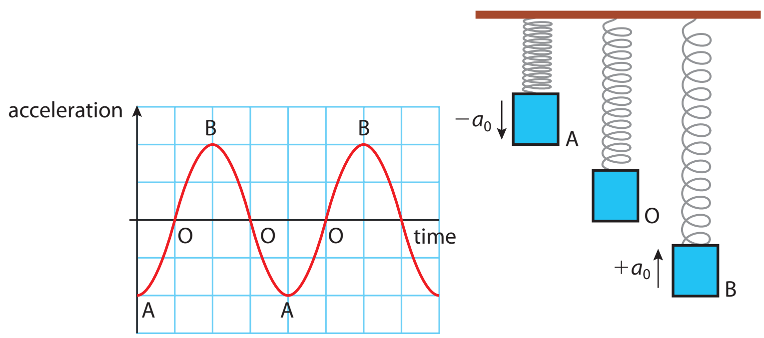


Figure 4.8 Acceleration–time graph.

The equation of this line is $a = -a_0 \cos \omega t$ where a_0 is this maximum acceleration.
So $x = x_0 \cos \omega t$ and $a = -a_0 \cos \omega t$

When displacement increases, acceleration increases proportionally but in a negative sense; in other words: $a \propto -x$

We have confirmed that the acceleration of the body is directly proportional to the displacement of the body and always directed towards a fixed point.

Worked example

A mass on a spring is oscillating with a frequency 0.2 Hz and amplitude 3.0 cm. What is the displacement of the mass 10.66 s after it is released from the top?

Solution

$x = x_0 \cos \omega t.$ Since this is SHM

where x = displacement

x_0 = amplitude = 3 cm

ω = angular velocity = $2\pi f = 2\pi \times 0.2$

$= 0.4\pi$ Hz

t = time = 10.66 s

$x = 0.03 \times \cos(0.4\pi \times 10.66)$ Substituting into the formula

$x = 0.02$ m

= 2 cm

Radians
When calculating cos ωt, you must have your calculator set on radians.

Exercises

3 For the same mass on a spring in the example above, calculate the displacement after 1.55 s.

4 Draw a displacement time sketch graph for this motion.

5 A long pendulum has time period 10 s. If the bob is displaced 2 m from the equilibrium position and released, how long will it take to move 1 m?

6 As a mass on a spring travels upwards through the equilibrium position, its velocity is 0.5 m s^{-1}. If the frequency of the pendulum is 1 Hz what will the velocity of the bob be after 0.5 s?

SHM and circular motion

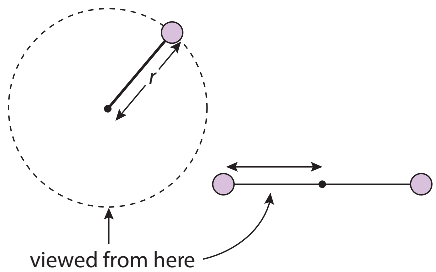


Figure 4.9 A short time after the ball starts moving, the radius makes an angle θ with the horizontal.

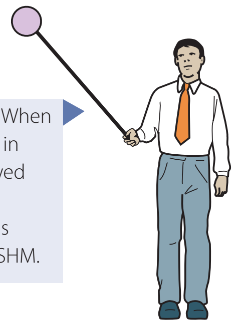


Figure 4.10 When a ball moving in a circle is viewed from the side, it looks like it is moving with SHM.

If we analyse the motion of the ball in Figure 4.9, we find that it is also SHM. The ball is travelling in a circle of radius x_0 with a constant speed v. The ball takes a time T to complete one revolution.

Let us consider the horizontal component of motion. We can write equations for this component.

Displacement

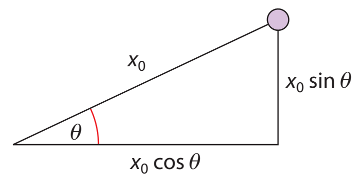


Figure 4.11 From the triangle we can see that *the horizontal displacement* $x = x_0 \cos\theta$.

Velocity

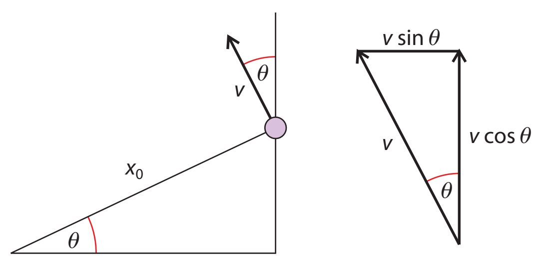


Figure 4.12 Horizontal velocity vectors.

Speed, $v = \dfrac{\text{distance}}{\text{time}}$

$= \dfrac{\text{circumference}}{\text{time period}}$

$= \dfrac{2\pi r}{T}$

But $\dfrac{2\pi}{T} = \omega$

So speed $= \omega r$

Centripetal acceleration $= \dfrac{v^2}{r}$

$= \dfrac{\omega^2 r^2}{r}$

$= \omega^2 r$

The horizontal velocity $= -v \sin\theta$

But for circular motion $v = \omega r$ so in this case $v = \omega x_0$

So horizontal velocity $= -\omega x_0 \sin\theta$

Acceleration

When bodies travel in a circle, they have an acceleration towards the centre (the centripetal acceleration) $a = \omega^2 r$. In this case, acceleration is $\omega^2 x_0$ since the radius is x_0.

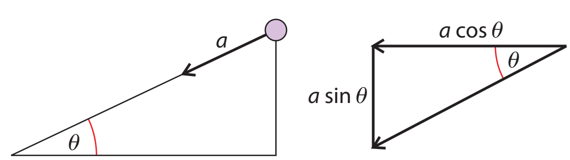


Figure 4.13 Horizontal acceleration vectors.

Horizontal component of acceleration $= -a \cos\theta$

But $a = \omega^2 x_0$

So horizontal acceleration $= -\omega^2 x_0 \cos\theta$

Now we have found that the displacement $x = x_0 \cos\theta$

So acceleration$= -\omega^2 x$

So the horizontal acceleration is proportional to the displacement, and is always directed towards the centre. In other words, the horizontal component of the motion is SHM. We have also found out that the constant of proportionality is ω^2.

Now we have concluded that this motion is SHM we can use the equations that we have derived to model all simple harmonic motions.

If you have done differentiation in maths then you will understand that if

displacement $x = x_0 \cos\omega t$

then velocity $\dfrac{dx}{dt} = -x_0\omega \sin\omega t$

and acceleration, $\dfrac{d^2x}{dt^2} = -x_0\omega^2 \cos\omega t$

This implies that $a = -\omega^2 x$

This is a much shorter way of deriving this result!

Equations for SHM

Displacement $= x_0 \cos \omega t$ (1)

Velocity $= -\omega x_0 \sin \omega t$ (2)

Acceleration $= -\omega^2 x_0 \cos \omega t$ (3)

We also know that $a = -\omega^2 x$

From Pythagoras, $1 = \sin^2 \theta + \cos^2 \theta$

So, $\sin \theta = \sqrt{1 - \cos^2 \theta}$ Rearranging

Therefore $\sin \omega t = \sqrt{1 - \cos^2 \omega t}$ Substituting for $\theta = \omega t$

Multiplying by ω gives $\omega x_0 \sin \omega t = \omega x_0 \sqrt{1 - \cos^2 \omega t}$

$\omega x_0 \sin \omega t = \omega \sqrt{x_0{}^2 - x_0{}^2 \cos^2 \omega t}$ Taking x_0 into the square root

But from equation (1) $x_0{}^2 \cos^2 \omega t = x^2$

So $v = \omega \sqrt{x_0{}^2 - x^2}$ Substituting

The maximum velocity is when the displacement is 0 so $x = 0$

Maximum velocity $= \omega x$

Worked example

1 A pendulum is swinging with a frequency of 0.5 Hz. What is the size and direction of the acceleration when the pendulum has a displacement of 2 cm to the right?

2 A pendulum bob is swinging with SHM at a frequency of 1Hz and amplitude 3 cm. At what position will the bob be moving with maximum velocity and what is the size of the velocity?

Solution

1 Assuming the pendulum is swinging with SHM, then we can use the equation $a = \omega^2 x$ to calculate the acceleration.

$\omega = 2\pi f = 2\pi \times 0.5 = \pi$

$a = -\pi^2 \times 0.02 = \mathbf{-0.197\ m\ s^{-2}}$ Since $-$ve direction is to the left

2 $v = \omega \sqrt{x_0{}^2 - x^2}$ Since the motion is SHM

This is maximum when $x = 0$ This is when the pendulum swings through the central position

The maximum value $= \omega x_0$ where $\omega = 2\pi f = 2 \times \pi \times 1 = 2\pi$ rad s^{-1}
Maximum $v = 2\pi \times 0.03 = \mathbf{0.188\ m\ s^{-1}}$

Summary
If a body oscillating with SHM has an angular frequency $\omega = 2\pi f$ and amplitude x_0 then its displacement (x), velocity (v) and acceleration (a) at any given time can be found from the following equations:
$x = x_0 \cos \omega t$
$v = -x_0 \omega \sin \omega t$
$a = -\omega^2 x_0 \cos \omega t$
In addition, at a given displacement x, the velocity and acceleration can be found from the following equations:
$v = \omega \sqrt{x_0{}^2 - x^2}$
Maximum velocity $= \omega x_0$
$a = -\omega^2 x$

The real pendulum
The pendulum is a classic example of SHM. However it is only SHM if the swings are very small (less than 10°). This is worth remembering if you ever do an experiment with a real pendulum.

Exercises

7 A long pendulum swings with a time period of 5 s and an amplitude of 2 m.
(a) What is the maximum velocity of the pendulum?
(b) What is the maximum acceleration of the pendulum?

8 A mass on a spring oscillates with amplitude 5 cm and frequency 2 Hz. The mass is released from its highest point. Calculate the velocity of the mass after it has travelled 1cm.

9 A body oscillates with SHM of time period 2 s. What is the amplitude of the oscillation if its velocity is 1 m s^{-1} as it passes through the equilibrium position?

4.2 Energy changes during simple harmonic motion (SHM)

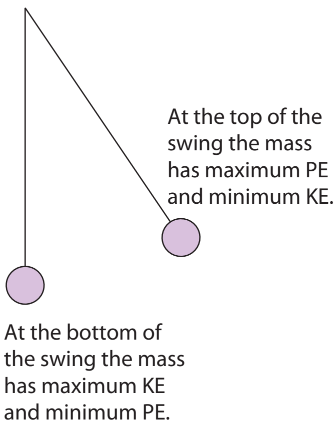
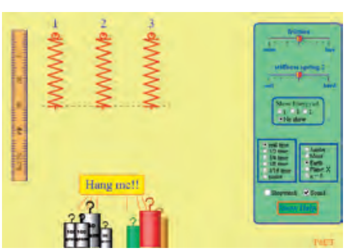


Figure 4.14 In the simple pendulum, energy is changing from one form to another as it moves.



To view the PhET *Masses and springs* simulation, visit heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 4.1.

Assessment statements

4.2.1 Describe the interchange between kinetic energy and potential energy during SHM.

4.2.2 Apply the expression $E_K = \frac{1}{2}m\omega^2(x_0^2 - x^2)$ for the kinetic energy of a particle undergoing SHM, $E_T = \frac{1}{2}m\omega^2 x_0^2$ for the total energy and $E_P = \frac{1}{2}m\omega^2 x^2$ for the potential energy.

4.2.3 Solve problems, both graphically and by calculation, involving energy changes during SHM.

If we once again consider the simple pendulum, we can see that its energy changes as it swings.

Kinetic energy

We have already shown that the velocity of the mass is given by the equation

$$v = \omega\sqrt{x_0^2 - x^2}$$

From definition, KE $= \frac{1}{2}mv^2$

Substituting: KE $= \frac{1}{2}m\omega^2(x_0^2 - x^2)$

KE is a maximum at the bottom of the swing where $x = 0$.

So $KE_{max} = \frac{1}{2}m\omega^2 x_0^2$

At this point the PE is zero.

Total energy

The total energy at any moment in time is given by:

total energy = KE + PE

So at the bottom of the swing:

total energy $= \frac{1}{2}m\omega^2 x_0^2 + 0 = \frac{1}{2}m\omega^2 x_0^2$

Since no work is done on the system, according to the law of conservation of energy, the total energy must be constant.

So total energy $= \frac{1}{2}m\omega^2 x_0^2$

Potential energy

Potential energy at any moment = total energy − KE

So PE $= \frac{1}{2}m\omega^2 x_0^2 - \frac{1}{2}m\omega^2(x_0^2 - x^2)$

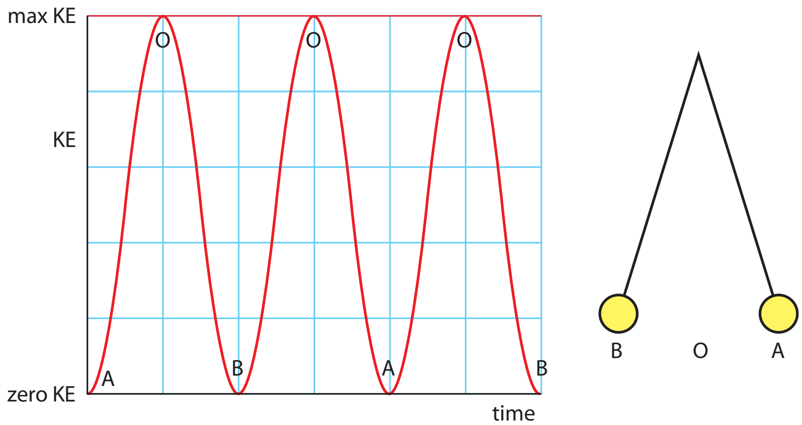
PE $= \frac{1}{2}m\omega^2 x^2$

Solving problems graphically

Kinetic energy

From previous examples we know that the velocity, $v = -v_0 \sin \omega t$

So $\frac{1}{2}mv^2 = \frac{1}{2}mv_0^2 \sin^2 \omega t$



The kinetic energy is a maximum when the bob is travelling fastest; this is at the bottom of the swing. At the top of the swing, the bob is stationary, so the KE is zero.

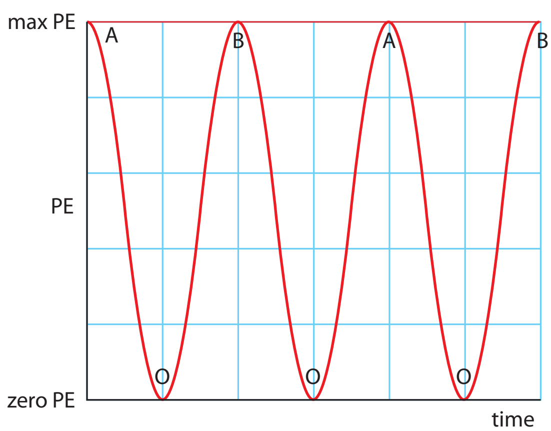
Figure 4.15 The graph of KE vs time is a $\sin^2$ curve.

Potential energy

The graph of PE can be found from $PE = \frac{1}{2} m\omega^2 x^2$

Since $x = x_0 \cos \omega t$

$PE \quad = \frac{1}{2}m\omega^2 x_0^2 \cos^2 \omega t = \frac{1}{2}mv_0^2 \cos^2 \omega t$

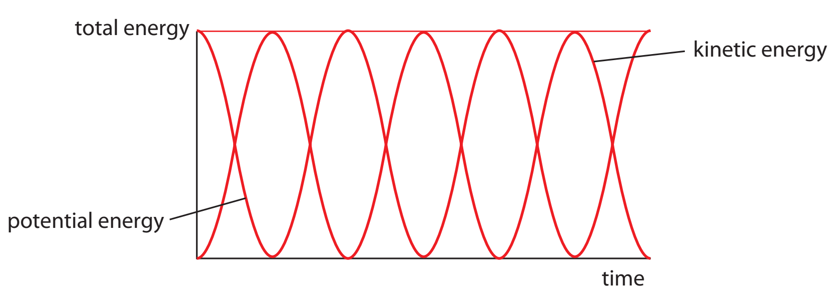


Potential energy
The potential energy is a minimum when the bob is at its lowest point; we take this to be zero. At the top of the swing, the potential energy is a maximum value.

Figure 4.16 The graph of PE vs time is a $\cos^2$ curve.

Total energy

If these two graphs are added together it gives a constant value, the total energy. (This might remind you of Pythagoras: $1 = \cos^2\theta + \sin^2\theta$.)



Total energy
If no energy is lost then the total energy is a constant value. When the bob is swinging, the energy continually changes between kinetic and potential.

Figure 4.17 Total energy vs time.

Worked example

1 A pendulum bob of mass 200 g is oscillating with amplitude 3 cm and frequency 0.5 Hz. How much KE will the bob have as it passes through the origin?

Solution

1 Since the bob has SHM, $KE_{max} = \frac{1}{2}m\omega^2 x_0^2$

where $x_0 = 0.03$ m and $\omega = 2\pi f = 2\pi \times 0.5 = \pi$

$KE_{max} = \frac{1}{2} \times 0.2 \times \pi^2 \times (0.03)^2 = \mathbf{8.9 \times 10^{-4}\,J}$

4.3 Forced oscillations and resonance

Assessment statements

4.3.1 State what is meant by *damping*.

4.3.2 Describe examples of damping.

4.3.3 State what is meant by *natural frequency of vibration* and *forced oscillations*.

4.3.4 Describe graphically the variation with forced frequency of the amplitude of vibration of an object close to its natural frequency of vibration.

4.3.5 State what is meant by *resonance*.

4.3.6 Describe examples of resonance where the effect is useful and where it should be avoided.

Damping

When deriving the equations for KE and PE, we assumed that no energy was lost. In real oscillating systems there is always friction and sometimes also air resistance. The system has to do work against these forces resulting in a loss of energy. This effect is called *damping*.

The suspension of a car. The damper is the red telescopic part.

A car suspension system has many springs between the body and the wheels. Their purpose is to absorb shock caused by bumps in the road.

The car is therefore an oscillating system that would oscillate up and down every time the car went over a bump. As this would be rather unpleasant for the passengers, the oscillations are damped by dampers (wrongly known as shock absorbers).

Light damping

If the opposing forces are small, the result is a gradual loss of total energy. This means that the amplitude of the motion gets slowly less with time. For example, a mass on a spring hanging in the air would have a little damping due to air resistance.

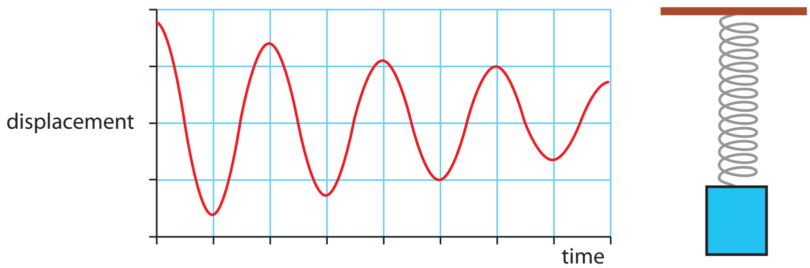


Figure 4.18 Reduction in amplitude due to light damping.

Frequency of damped harmonic motion
You can see from the graph that the frequency does not change as the amplitude gets less. As the motion slows down, the distance travelled gets less, so the time for each cycle remains the same.

If the mass is suspended in water, the damping is greater, resulting in a more rapid loss of energy.

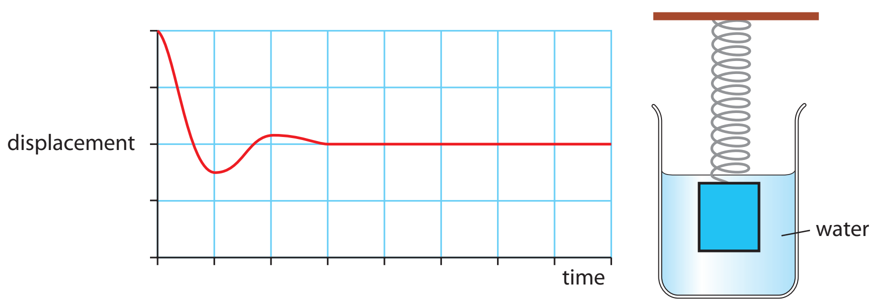


Figure 4.19 Reduction in amplitude due to heavier damping.

Critical damping

Critical damping occurs if the resistive force is so big that the system returns to its equilibrium position without passing through it. This would be the case if the mass were suspended in a thicker liquid such as oil.

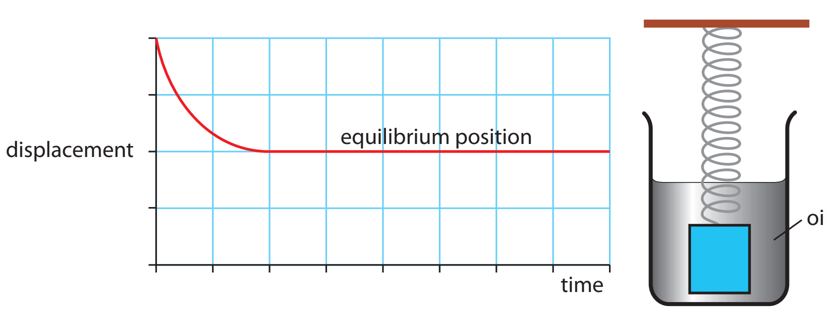


Figure 4.20 Reduction in amplitude due to critical damping.

Resonance

It's possible to shatter a wine glass if you sing at its natural frequency.

In all of the previous examples, a system has been displaced and released, causing an oscillation. The frequency of this oscillation is called the *natural frequency*. If a system is forced to oscillate at a frequency other than the natural frequency, this is called a *forced oscillation*.

Resonance is an increase in amplitude that occurs when an oscillating system is forced to oscillate at its own natural frequency.

For example, when you hit a wine glass with your finger, it vibrates. If you sing at the same frequency, your voice can cause the wine glass to resonate. Sing loud enough and the wine glass will shatter (not many people can do this).

If a spring is pulled down and released, then it will oscillate at its own natural frequency. If the support is oscillated, then the system will be forced to vibrate at another frequency. If the driving frequency is the same as the natural frequency, then resonance occurs.

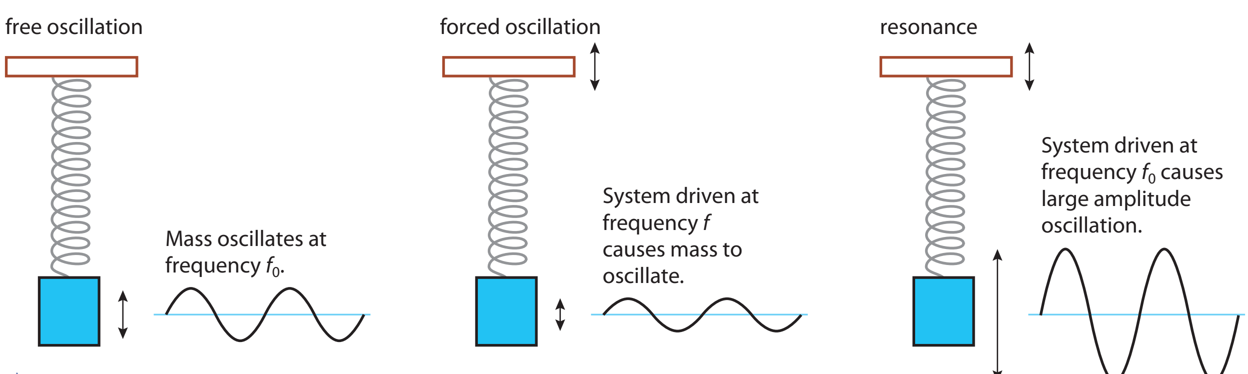


Figure 4.21 The effect of varying driving frequency.

The radio tuner
When you tune your radio, you are adjusting an electric circuit so that it resonates with the signal of a particular frequency. If the resonance curve for the circuit were not sharp, you would be able to tune into the station over a wide range of frequencies, and would be likely to get interference from other stations.

Resonance curve

A graph of the amplitude of oscillation against the driving frequency is called a resonance curve. The sharpness of the peak is affected by the amount of damping in the system.

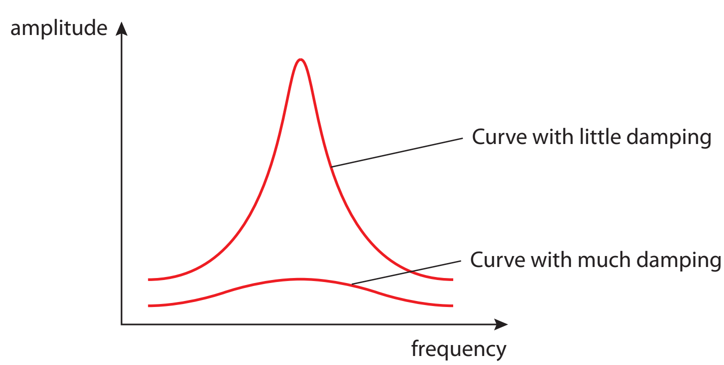


Figure 4.22 Graph of amplitude vs frequency for different levels of damping.

Phase

If we take two identical pendulum bobs, displace each bob to the right and release them at the same time, then each will have the same displacement at the same time. We say the oscillations are *in phase*. If one is pulled to the left and the other to the right, then they are *out of phase*.

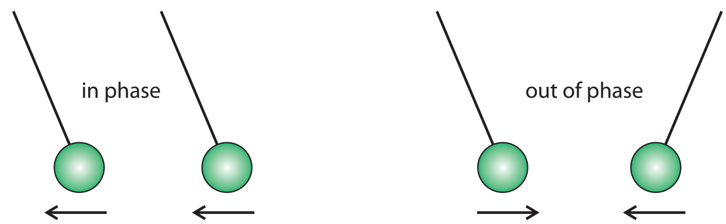


Figure 4.23 The pendulum bobs are in phase when they swing together.

This can be represented graphically:

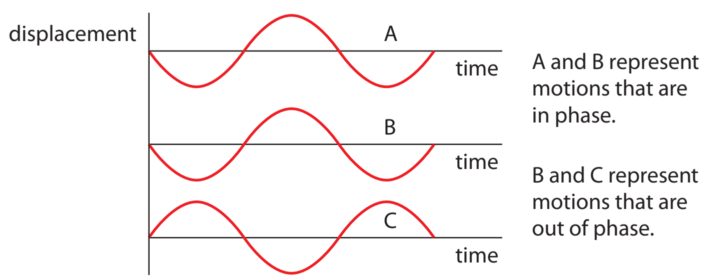


Figure 4.24 Displacement–time graphs for bodies in and out of phase.

Riding a horse
When riding a horse it is important to stay in phase with the horse. If you are out of phase, then you will be coming down when the horse is going up, resulting in an uncomfortable experience. If the horse goes up and down too fast, then it can be very difficult to stay in phase. A mechanical horse is more difficult to ride; you can only accelerate downwards at 9.8 m s^{-2}, so if the horse accelerates down too fast then you can't keep up with it.

Phase difference

The phase difference is represented by an angle (usually in radians). We can see from the previous graphs that if two oscillations are completely out of phase then the graphs are displaced by an angle π. We say the *phase difference* is π.

When juggling balls (or oranges) they go up and down at different times – they are out of phase.

Worked example

A ball is sitting on a platform oscillating with amplitude 1 cm at a frequency of 1 Hz. As the frequency is increased, the ball starts to lose contact with the platform. At what frequency does this take place?

Solution

The ball will lose contact when the acceleration of the platform is greater than 9.8 m s^{-2}.

Using the formula $a = -\omega^2 x_0$

$$\omega = \sqrt{9.8/0.01} = 31.3 \text{ rads s}^{-1}$$

$$f = \frac{\omega}{2\pi} = \mathbf{5\ Hz}$$

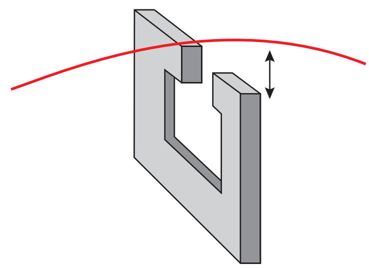
Experimental measurement of oscillations

The frequency at which a child oscillates on a swing is low enough to measure using a stopwatch (although to be accurate, you should use some sort of marker, so you can easily judge when the child has made a complete cycle). Higher frequency oscillations are not possible to measure in this way but can be measured using electronic sensors. Here are some examples of how you could make those measurements.

Photo gate

A photo gate sends a signal to a computer each time something passes through it. If a vibrating object passes through the gate each cycle then the computer can calculate the time period of the oscillation.

Figure 4.25 A photo gate is used to measure the time period of a vibrating elastic string.

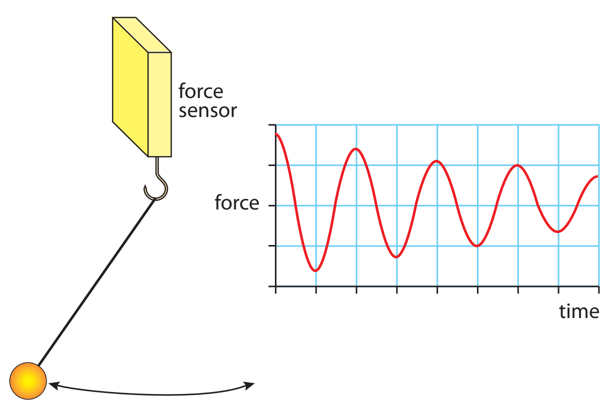


The computer will record each time the string passes through the gate. The time period is the time between the first pass and the third pass. Depending on the software used, it may be possible for the computer to calculate and display the frequency.

Force sensor

When a pendulum swings, the tension in the string varies with time. A force sensor can be used to measure the tension, enabling you to plot a graph of tension *vs* time on the computer. The frequency is calculated from the graph.

Figure 4.26 The maximum force gets less as the amplitude gets less.



With this method it is also possible to see the damping of the motion.

Position sensor

To measure an oscillation using a position sensor, the oscillating body must move backwards and forwards (or up and down) in front of the sensor. The sensor sends out a sound that is reflected off the object back to the sensor. By measuring the time taken for the sound to reflect back from the object, the computer can calculate the distance between the sensor and the object. This method has the advantage of not disturbing the motion, but the object must be big enough for the sensor to detect it.

4.4 Wave characteristics

Assessment statements

4.4.1 Describe a wave pulse and a continuous progressive (travelling) wave.
4.4.2 State that progressive (travelling) waves transfer energy.
4.4.3 Describe and give examples of transverse and of longitudinal waves.
4.4.4 Describe waves in two dimensions, including the concepts of wavefronts and of rays.
4.4.5 Describe the terms *crest*, *trough*, *compression* and *rarefaction*.
4.4.6 Define the terms *displacement*, *amplitude*, *frequency*, *period*, *wavelength*, *wave speed* and *intensity*.
4.4.7 Draw and explain displacement–time graphs and displacement – position graphs for transverse and for longitudinal waves.
4.4.8 Derive and apply the relationship between wave speed, wavelength and frequency.
4.4.9 State that all electromagnetic waves travel with the same speed in free space, and recall the orders of magnitude of the wavelengths of the principal radiations in the electromagnetic spectrum.

The word *wave* was originally used to describe the way that a water surface behaves when it is disturbed. We use the same model to explain sound, light and many other physical phenomena. This is because they have some similar properties to water waves, so let's first examine the way water waves spread out.

If a stone is thrown into a pool of water, it disturbs the surface. The disturbance spreads out or *propagates* across the surface, and this disturbance is called a *wave*. Observing water waves, we can see that they have certain basic properties (in other words, they do certain things).

Reflection

If a water wave hits a wall, the waves reflect.

Sea waves reflect off a cliff.

Waves change direction as they approach a beach.

Refraction

When sea waves approach a beach, they change direction because of the difference in height of different parts of the sea floor. This causes the waves to bend.

Interference

When two waves cross each other, they can add together creating an extra big wave.

Diffraction

When water waves pass through a small opening, the waves spread out.

Anything that reflects, refracts, interferes and diffracts can also be called a wave.

To view the PhET *Waves on a string* simulation, visit heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 4.2.

One-dimensional waves

The next step is to derive a model for wave motion and use it to help us understand why waves behave in the way that they do. However, since water waves are two-dimensional, they are not the easiest waves to start with. We will begin by looking at two examples of one-dimensional waves: waves in a string and waves in a spring.

Wave pulse in a string

If a string held between two people is displaced (flicked), a disturbance can be seen to travel from one end to the other. This is called a *wave pulse.*

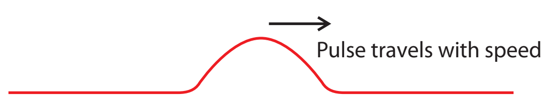


Figure 4.27 A wave pulse.

We can see that the pulse travels with a certain speed – this is called the *wave speed.*

Wave speed is the distance travelled by the wave profile per unit time.

Note: No part of the string actually moves in the direction of the wave velocity – in fact, each particle in the string moves at right angles to the direction of wave velocity.

Reflection of a wave pulse

If the pulse meets a fixed end (e.g. a wall), it exerts an upward force on the wall. The wall being pushed up, pushes back down on the string sending an inverted reflected pulse back along the string.

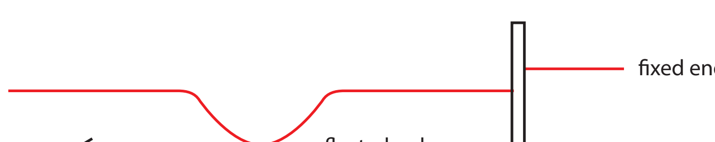


Figure 4.28 A reflected pulse.

Interference of wave pulses

If two pulses are sent along a string from each end, they will cross each other in the middle of a string.

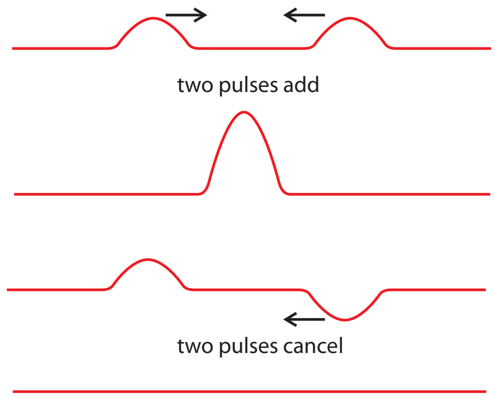


Figure 4.29 The resultant wave is the sum of the individual waves.

Since waves in a string do not spread out, they cannot diffract or refract. You would have to observe the 2D equivalent, waves in a rubber sheet, to see this.

Transfer of energy

It can be seen that as the string is lifted up it is given PE. This PE is transferred along the string. A wave can therefore be thought of as a transfer of energy. There is in fact so much energy transferred by waves in the sea that they can be used to produce electricity.

Continuous waves in a string

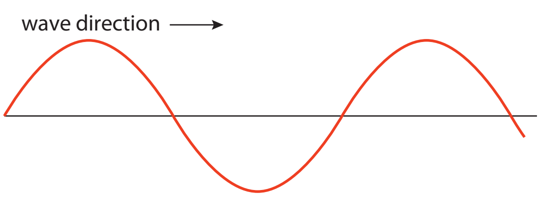


Figure 4.30 The 'sine shape' or profile moves along the string with the wave speed.

If the end of a string is moved up and down with simple harmonic motion of frequency *f*, a series of pulses moves along the string in the shape of a sine curve, as in Figure 4.30.

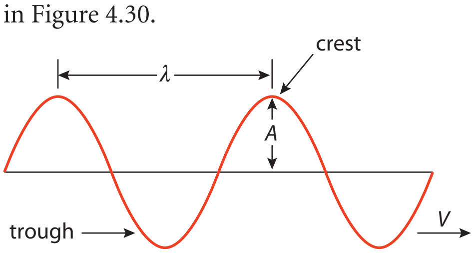


Figure 4.31 The quantities used to define a wave.

Amplitude (*A*)
The maximum displacement of the string from the equilibrium position.

Wave speed (*v*)
The distance travelled by the wave profile per unit time.

Wavelength (*λ*)
The distance between two consecutive crests or any two consecutive points that are in phase.

Frequency (*f*)
The number of complete cycles that pass a point per unit time.

Period (*T*)
Time taken for one complete wave to pass a fixed point ($T = 1/f$)

Phase
The phase is a quantity that tells us whether parts of a wave go up and down at the same time or not.

Transverse waves

Figure 4.32 Transverse wave

A wave in a string is an example of a transverse wave. The direction of disturbance is perpendicular to the direction that the wave profile moves.

Stringed instruments

When you pluck the string of a guitar, a wave reflects backwards and forwards along the string. The vibrating string creates the sound that you hear. The pitch of the note is related to the frequency of the string (high pitch = high frequency).

- Why are the low notes thick strings?
  The speed of the wave is inversely related to the mass per unit length of the string. Thick strings have a greater mass per unit length, so the wave will travel more slowly in a thick string. If we rearrange the formula $v = f\lambda$, we find that $f = \frac{v}{\lambda}$ so reducing the wave speed will reduce the frequency of the wave.
- Why does shortening the string make the note higher?
  Shortening the string reduces the wavelength of the wave. According to the formula $f = \frac{v}{\lambda}$, reducing the wavelength will increase the frequency.
- Why does tightening the string make the note higher?
  The wave speed is directly related to the tension in the string. Increasing tension increases the wave speed, which, according to the formula $f = \frac{v}{\lambda}$, will increase the frequency of the wave.

Worked example

1 The A string of a guitar vibrates at 110 Hz. If the wavelength is 153 cm, what is the velocity of the wave in the string?

2 A wave in the ocean has a period of 10 s and a wavelength of 200 m. What is the wave speed?

Relationship between f and λ

$v = f\lambda$

If the frequency is f then the time for the wave to progress one cycle is $1/f$. In this time the wave has moved forward a distance equal to one wavelength (λ).

$$\text{Velocity} = \frac{\text{distance}}{\text{time}}$$

$$v = \frac{\lambda}{1/f} = f\lambda$$

Solution

1 $v = f\lambda$

$f = 110\ \text{Hz}$ and $\lambda = 1.53\ \text{m}$

• **Examiner's hint:** Change cm to m.

$v = 110 \times 1.53\ \text{m s}^{-1}$

$= \mathbf{168.3\ m\ s^{-1}}$

2 $T = 10\ \text{s}$

$f = 1/T\ \text{Hz}$

$= 0.1\ \text{Hz}$

$v = f\lambda$

$v = 0.1 \times 200\ \text{m s}^{-1}$

$= \mathbf{20\ m\ s^{-1}}$

Waves in a spring

If a long soft spring (a slinky) is stretched and one end moved back and forth, a compression can be seen to travel along it. Although this may not look like a wave, it *is* transferring energy from one end to the other and so fits the definition.

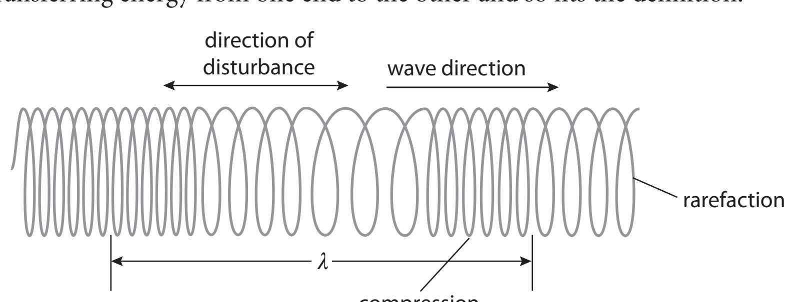


Figure 4.33 The difference between a compression wave in a spring and the transverse wave in a string is the direction of disturbance.

Longitudinal waves

Figure 4.34 Longitudinal wave.

A compression wave in a slinky is an example of a longitudinal wave. In a longitudinal wave, the disturbance is parallel to the direction of the wave.

Reflection

When the wave in a spring meets a fixed end, it will reflect.

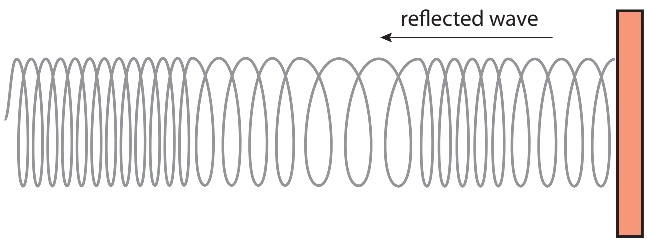


Figure 4.35 A wave in a spring is reflected off a wall.

Earthquake waves
An earthquake is caused when parts of the Earth's crust move against each other. This disturbance causes both longitudinal and transverse waves to spread around the Earth.

Transverse wave
When an earthquake occurs the ground shakes up and down.

Longitudinal wave
The movement in the Earth's crust compresses the rock.

Interference

Although not easy to observe, when two longitudinal waves meet, the displacements superpose in the same way as transverse waves.

Light and sound
Both light and sound are disturbances that spread out, so can be thought of as waves. Light can be polarized (for example, by Polaroid sunglasses) but sound cannot. This is one way to tell that light is transverse and sound is longitudinal.

Distinguishing longitudinal and transverse

A wave is polarized if the displacement is only in one direction.

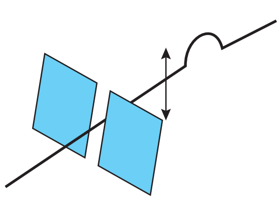


Figure 4.36 A string wave can be polarized by passing through a narrow slit.

The string can only move up and down so a wave in this string will be polarized. To test if the wave is polarized we can place another slit on the string; the wave only passes if the slits are parallel. Only transverse waves can be polarized, so this property can be used to tell if a wave is transverse or longitudinal.

Graphical representation of a wave

There are two ways we can represent a wave graphically, either by drawing a displacement–time graph for one point on the wave, or a displacement–position graph for each point along the wave.

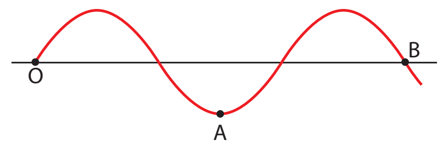


Figure 4.37 A snapshot of a transverse wave.

Displacement–time

Consider point A on the transverse wave in Figure 4.37.

Point A is moving up and down with SHM as the wave passes. At present, it is at a minimum of displacement. As the wave progresses past A, this point will move up and then down.

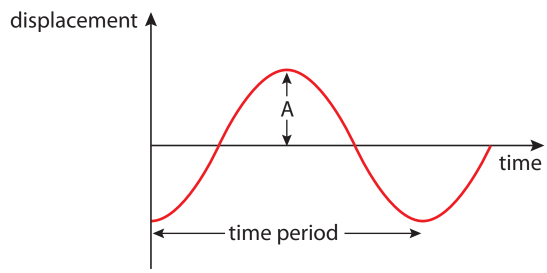


Figure 4.38 The displacement– time graph for point A.

Note: Because the horizontal axis is time, the separation of the peaks represents the time period, not the wavelength.

We can also draw a graph for point B. This point starts with zero displacement then goes up.

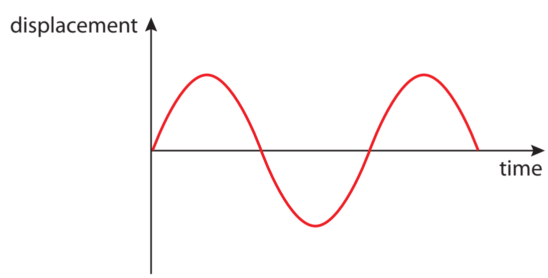


Figure 4.39 The displacement–time graph for point B.

Note: The event that will happen next is to the right on the graph but the part of the wave that will arrive next is to the left on the wave.

Displacement–position

To draw a displacement–position graph, we must measure the displacement of all the points on the wave at one moment in time.

Figure 4.40 shows the graph at the same time as the snapshot in Figure 4.37 was taken. The position is measured from point O.

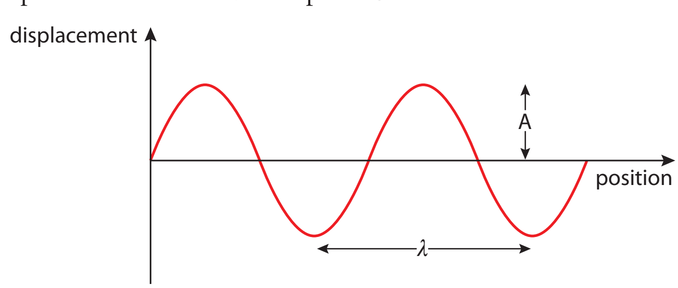


Figure 4.40 The displacement–position graph for all points at one time.

This is just like a snapshot of the wave – however, depending on the scale of the axis, it might not look quite like the wave.

Longitudinal waves

We can also draw graphs for a longitudinal wave. Consider a chain of balls connected with springs. If the red ball on the left were moved back and forth with SHM, it would send a longitudinal wave along the chain.

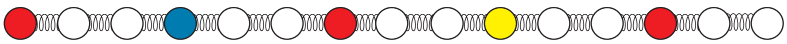


Figure 4.41 A line of balls joined by springs.

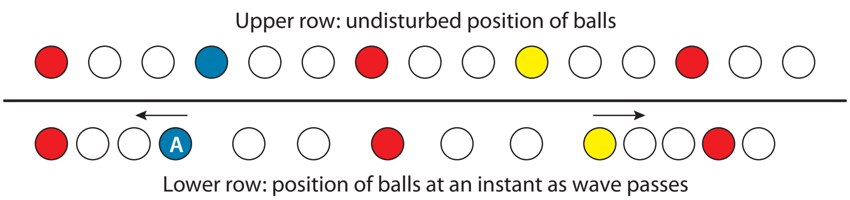


Figure 4.42 A snapshot taken as a wave passes through the chain.

Each ball simply moves back and forth with SHM. Ball A is at present displaced to the left. This ball has negative displacement.

Displacement–time

We can draw a displacement–time graph for ball A starting at the time of the snapshot in Figure 4.42.

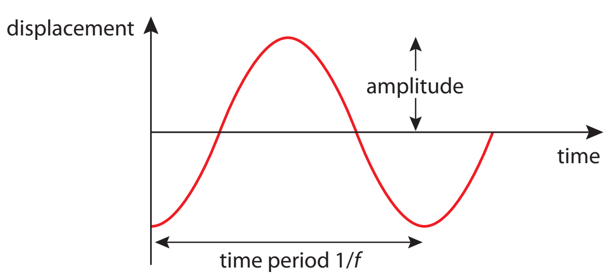


Figure 4.43 The displacement–time graph for point A.

Displacement–position

To draw a displacement–position graph, we must compare the position of each ball with its original position.

The red balls have not moved, so their displacement is 0.

The blue balls have moved to the left, so their displacement is negative.

The yellow ball has moved to the right, so its displacement is positive.

To see an animated version of this, visit heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 4.3.

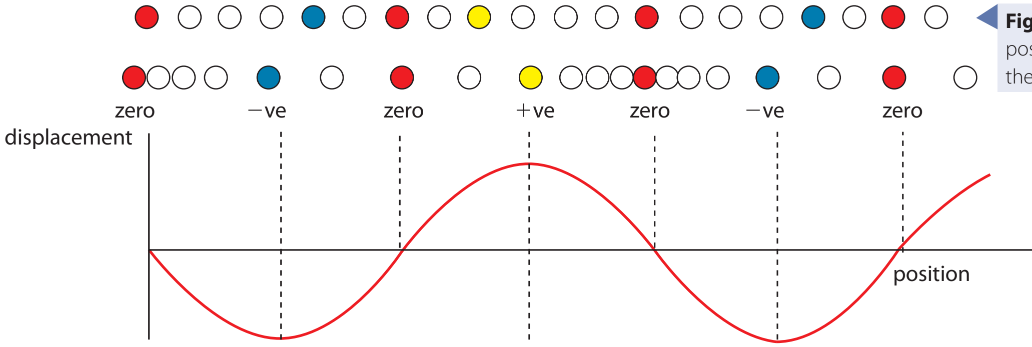
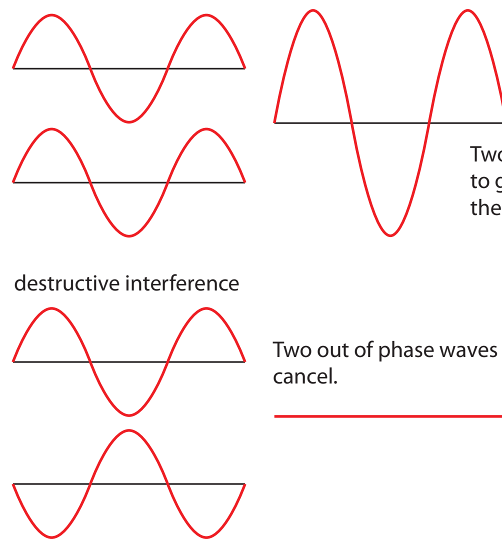


Figure 4.44 The displacement position graph for all points on the wave.

Superposition of one-dimensional waves

When two waves are incident along the same string, we can find the resultant wave by adding the individual displacements.

Figure 4.45 Superposition of waves.

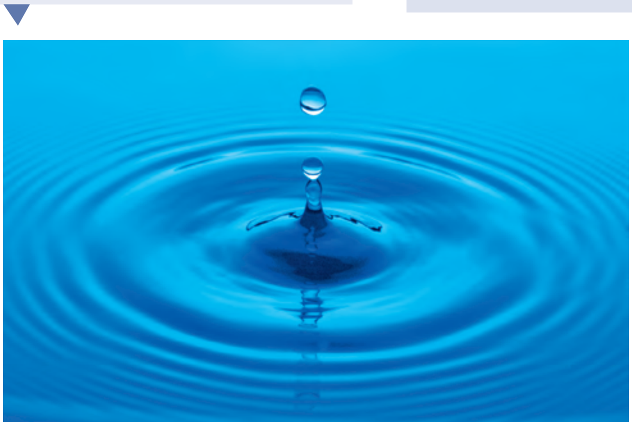


4.5 Wave properties

Assessment statements

4.5.1 Describe the reflection and transmission of waves at a boundary between two media.
4.5.2 State and apply Snell's law.
4.5.3 Explain and discuss qualitatively the diffraction of waves at apertures and obstacles.
4.5.4 Describe examples of diffraction.
4.5.5 State the principle of superposition and explain what is meant by constructive interference and by destructive interference.
4.5.6 State and apply the conditions for constructive and for destructive interference in terms of path difference and phase difference.
4.5.7 Apply the principle of superposition to determine the resultant of two waves

Ripples spreading out in a circle after the surface is disturbed.



Two-dimensional waves

We will now use water waves to model the motion of waves in 2D. If a disturbance is made by a point object in the middle of a tank of water, ripples spread out in circles across the tank. We will use pictures from a computer simulation to show the effect more clearly.

Wavefront

This is a line joining points that are in phase. The straight or circular lines that you can see in the photos are wavefronts.

Rays

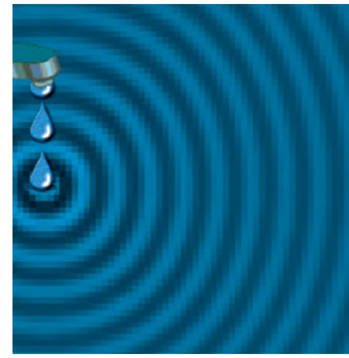
Rays are lines drawn to show the direction of the waves – they are always at right angles to the wavefront.

Circular wavefronts

A circular wavefront is produced by a point disturbance. The rays are radial, as they are perpendicular to the wavefronts.

Plane wavefront

Plane wavefronts are produced by an extended disturbance e.g. a long piece of wood dipped into the water, or a point that is so far away that the circles it produces look like straight lines.



To view the PhET simulation *Wave interference*, visit heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 4.4.

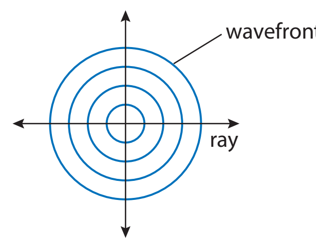


Figure 4.46 A circular wavefront spreading out from a point.

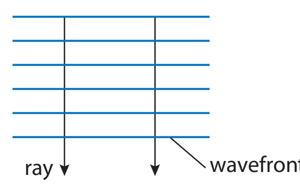


Figure 4.47 Parallel plane wavefronts.

A plane wavefront moves towards the beach.

Reflection

When a wave hits a barrier, it is reflected.

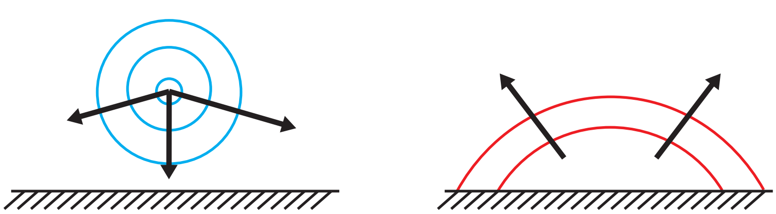


Figure 4.48 Reflection of a circular wavefront.

Notice how the reflected wave appears to originate from somewhere on the other side of the barrier. This is just the same as the appearance of an image of yourself behind a mirror.

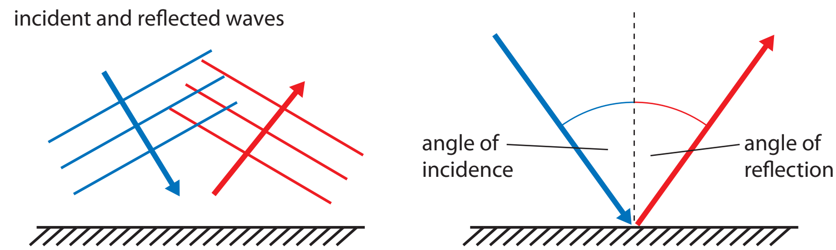
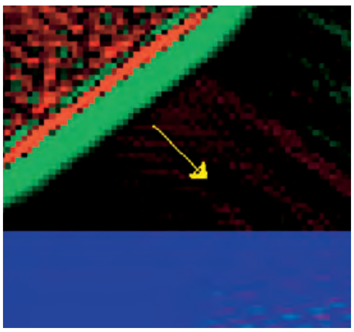
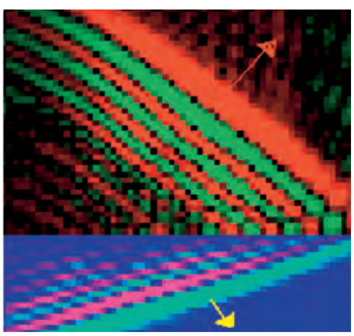


Figure 4.49 A plane wavefront is reflected at the same angle that it comes in at.





Rather than measuring the angle that the wavefront makes, it is more convenient to measure the angles that the rays make with a line drawn at 90° to the barrier. This line is called the *normal*.

The laws of reflection

The laws of reflection describe how waves are reflected from barriers.

- The angle of incidence = the angle of reflection.
- The incident and reflected rays are in the same plane as the normal.

Change of medium

Whenever a wave travels from one medium to another, part of the wave is reflected and part transmitted. An example of this is when light hits a glass window: most passes through but a fraction is reflected. So you see a reflection of yourself in the window and someone standing on the other side of the window can see you. The part of the wave that passes through the window is called the *transmitted* part.

To view screenshots showing the reflected and transmitted wave, visit heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 4.5.

Refraction

When a wave passes from one medium to another, its velocity changes. For example, when a water wave passes from deep water into shallow water, it slows down. If the wave hits the boundary between the media at an angle, then the wave also changes direction.

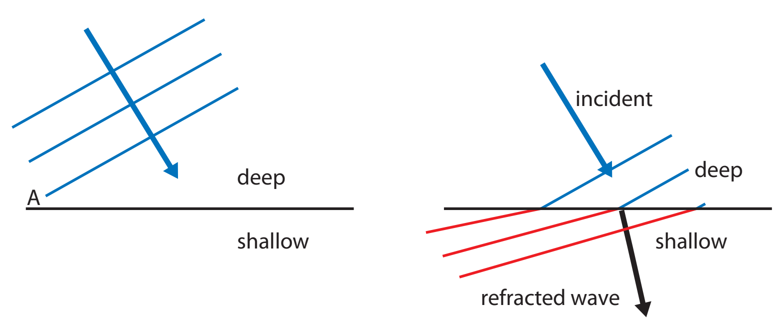


Figure 4.50 Refraction is the change of direction when a wave passes from one medium to another.

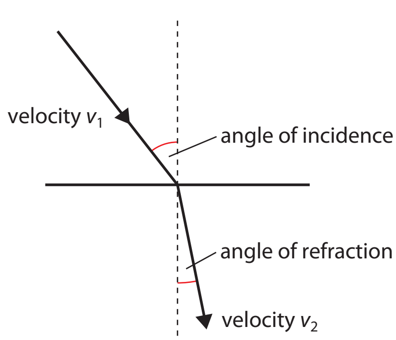


Figure 4.51 Angles of incidence and refraction.

Point A on the incident wave hits the boundary first, so this part of the wave then progresses into the shallow water more slowly. The rest of the wave in the deep water is still travelling fast so catches up with the slow moving part, causing the wavefront to change direction. This is simpler to see if we just draw the rays.

Snell's law

Snell's law relates the angles of incidence and refraction to the ratio of the velocity of the wave in the different media. The ratio of the sine of the angle of incidence to the sine of the angle of refraction is equal to the ratio of the velocities of the wave in the different media.

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2}$$

As can be seen from the example of the bent straw in the photo, light refracts when it passes from one medium to another. The ratio of the velocity of light in the two media is called the *refractive index*.

Light reflected off the straw is refracted as it comes out of the water causing the straw to appear bent.

Worked example

A water wave travelling at 20 m s^{-1} enters a shallow region where its velocity is 15 m s^{-1} (Figure 4.52). If the angle of incidence of the water wave to the shallow region is 50°, what is the angle of refraction?

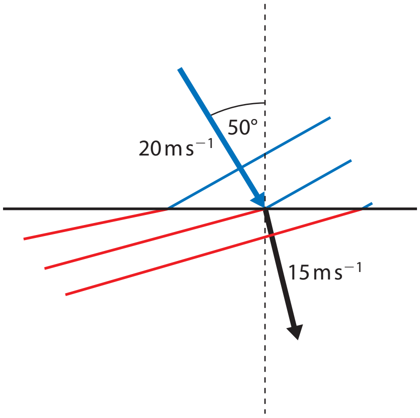


Figure 4.52 Always draw a diagram.

Solution

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2} = \frac{20}{15}$$

so $\sin r = \dfrac{\sin 50°}{1.33} = 0.576$ Applying Snell's law

$\boldsymbol{r = 35.2°}$

Refractive index

When light travelling in air is refracted by an optical medium, the ratio sin *i*/sin *r* is called the refractive index of the medium. If the refractive index is large it means that the light is refracted by a large angle.

Material	Refractive index
Water	1.33
Glass	1.50
Diamond	2.42

Exercises

Use the refractive indices in the table to solve the following problems.

10 Light travelling through the air is incident on the surface of a pool of water at an angle of 40°. Calculate the angle of refraction.

11 Calculate the angle of refraction if a beam of light is incident on the surface of a diamond at an angle of 40°.

12 If the velocity of light in air is 3×10^8 m s^{-1}, calculate its velocity in glass.

Diffraction

Diffraction takes place when a wave passes through a small opening. If the opening is very small, then the wave behaves just like a point source as shown below.

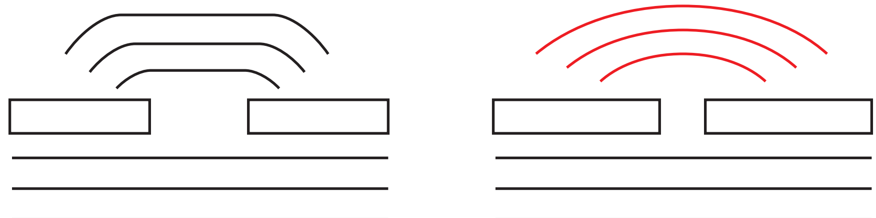
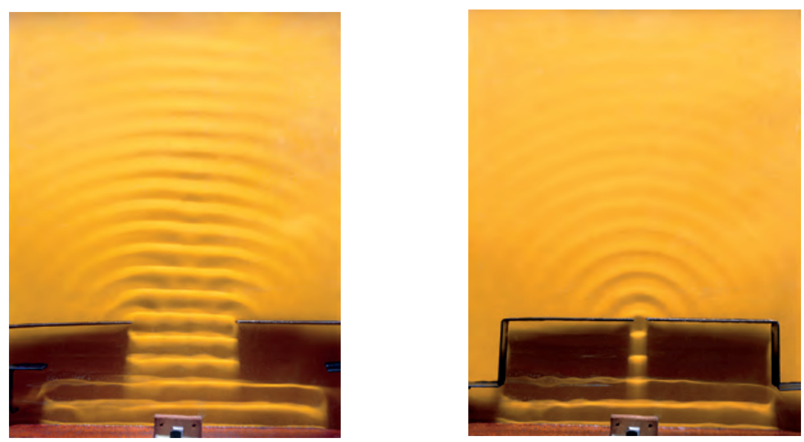


Figure 4.53 If the opening is a bit bigger then the effect is not so great.



Water waves diffracting through two different sized openings. The waves are diffracted more through the narrower opening.

Interference

When two-dimensional waves interfere, the phase difference between the two waves is different in different places. This means that in some places the waves add and in other places they cancel. This can be seen in the picture below. This shows two sources producing waves of the same frequency.

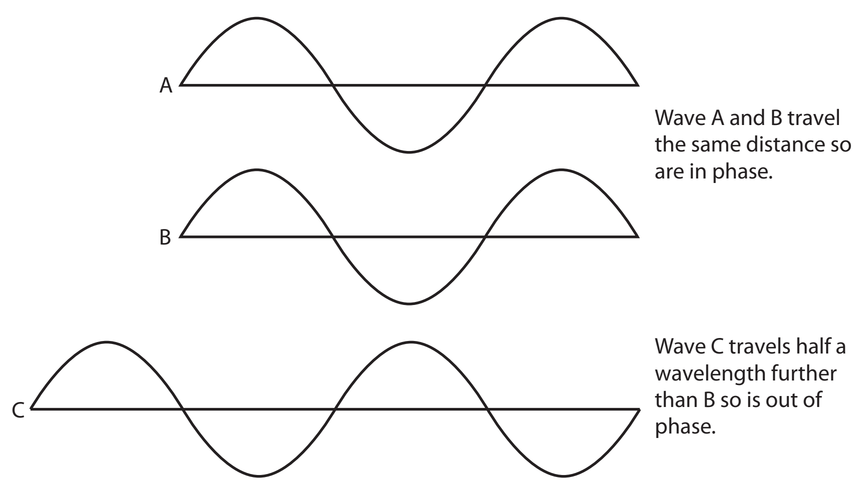


Figure 4.54 How phase difference is related to path difference.

If the path difference is a whole number of wavelengths, then the waves are *in phase*.

If the path difference is an odd number of half wavelengths then the waves are *out of phase*.

The effect of interference in two dimensions can be seen in Figure 4.55.

Identical waves from A and B spread out across the surface. At X, the waves from A and B have travelled the same distance, so are in phase and add together. At Y, the wave from B has travelled half a wavelength more then the wave from A, so the waves are out of phase and cancel out.

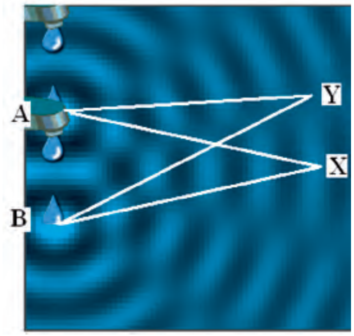


Figure 4.55 Interference effects seen in the PhET simulation. To view this, visit heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 4.6.

Phase angle

If the waves are completely out of phase then phase angle $= \pi$.

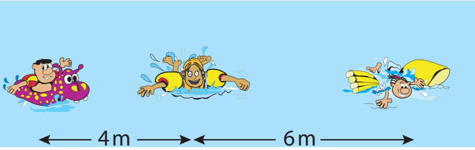
If out of phase but not completely out of phase, then the phase angle can be calculated from the path difference.

If path difference $= d$ then phase angle $\varphi = \dfrac{2\pi d}{\lambda}$

Worked example

Two boys playing in a pool make identical waves that travel towards each other. The boys are 10 m apart and the waves have a wavelength 2 m. Their little sister is swimming from one boy to the other. When she is 4 m from the first boy, will she be in a big wave or a small wave?

Figure 4.56 A diagram always helps, no matter how silly it is.



Solution

The waves from the boys will interfere when they meet, if the girl is 4 m from the first boy, then she must be 6 m from the other. This is a path difference of 2 m, one whole wavelength. The waves are therefore in phase and will add.

Exercise

13 Two wave sources A and B produce waves of wavelength 2 cm. What is the phase angle between the waves at

(a) a point C distance 6 cm from A and 6.2 cm from B?

(b) a point D distance 8 cm from A and 7 cm from B?

(c) a point E distance 10 cm from A and 11.5 cm from B?

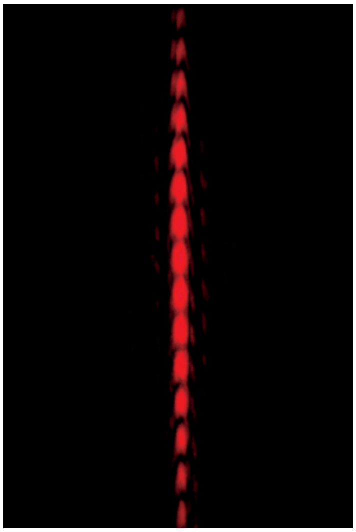
Examples of waves

Light

It is worth having a more detailed look at the wave properties of light. We have seen examples of how light reflects and refracts, and if light is a wave, then it must also interfere and diffract.

- Diffraction of light

We have seen that if a wave passes through an opening that is about the same size as its wavelength then it will spread out. If a beam of light passes through a narrow slit (close to the wavelength of light – around 500 nm), it will spread out.



The combined effect of diffraction and interference causes this pattern of dots when laser light passes through a pair of narrow slits.

- Interference of light

Waves only interfere if they have the same frequency and similar amplitude. If we take a source of light and split it into two we can create two identical (or *coherent*) wave sources. If the waves from these sources overlap, then areas of bright and dark are created, where the waves interfere constructively and destructively.

- Polarization

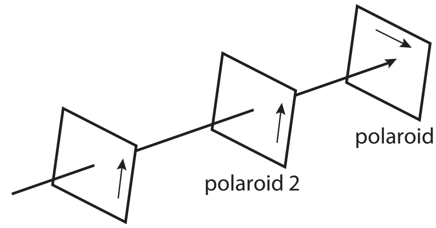
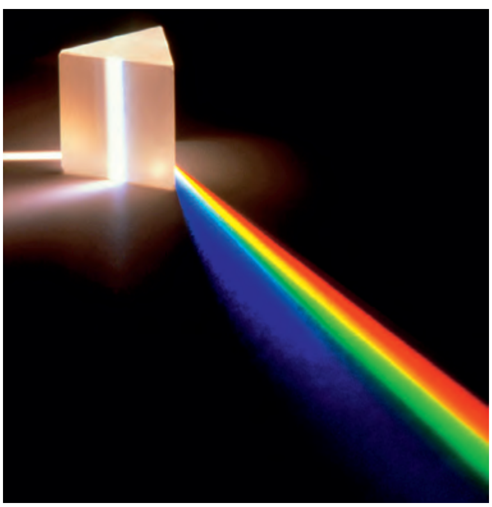


Figure 4.57 The light passes through polaroids 1 and 2, which have the same alignment, but not 3.

When light passes through polaroid sunglasses, it becomes polarized in one direction. We can test to see if the light is polarized by taking a second piece of polaroid and rotating it in front of the sunglasses. As we rotate the polaroid we find that the polarized light can only pass when the second piece is in the same orientation as the first.

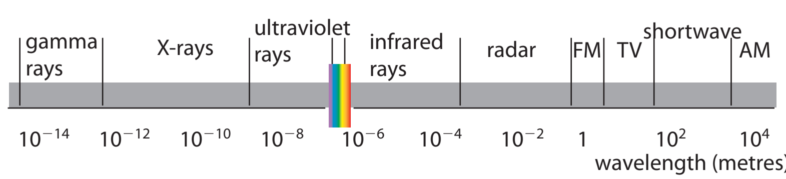


White light can be split up into its component colours by passing it through a prism.

- Wavelength and amplitude of light

Light is an electromagnetic (EM) wave; that is a propagation of transverse disturbance in an electric and magnetic field. Unlike the other types of waves considered here, EM waves can travel through a vacuum. As with all waves, light waves have wavelength and amplitude. The wavelength of light can vary from 400 nm to 800 nm, with different wavelengths having different colours. White light is made up of all colours mixed together, but if white light is passed through a prism, the different colours are split up, forming a spectrum. This happens because each wavelength refracts by a different amount, and therefore a different angle. This is what happens when a rainbow is formed.

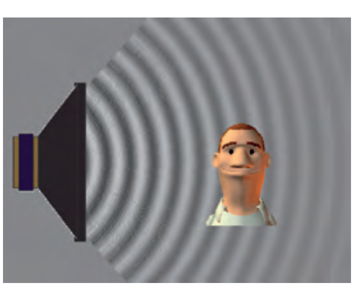
Visible light is just one small part of the complete EM spectrum. The full range of wavelength is from 10^{-14}m to 10^4m. Each part of the spectrum has different properties and a different name, as illustrated in the diagram below.



When we say light is a wave we mean it has the same properties as a wave. Does this mean it actually *is* a wave?

The amplitude of light is related to its brightness. The brightness of light is how we perceive light. The physical quantity that measures it is the light *intensity.* This is proportional to the square of the amplitude.

The speed of EM waves in a vacuum is $2.99 \times 10^8\ \mathrm{m\,s^{-1}}$.



To view the Phet sound waves simulation, visit heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 4.7.

Sound

- Reflection

If you shout in front of a cliff, the sound reflects back as an echo. In fact any wall is a good reflector of sound, so when speaking in a room, the sound is reflected off all the surfaces This is why your voice sounds different in a room and outside.

- Refraction

When sound passes from warm air into cold air, it refracts. This is why sounds carry well on a still night.

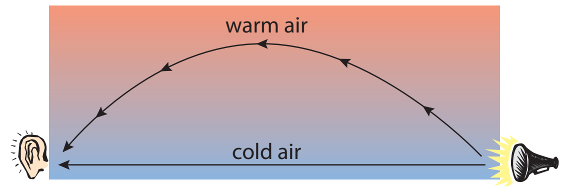


Figure 4.58 Sound refracts through layers of air.

The sound travels to the listener by two paths, one direct and one by refraction through the layers of air. This results in an extra loud sound.

- Diffraction and interference

Because sound reflects so well off all surfaces, it is very difficult to do sound experiments in the laboratory. This makes it difficult to observe sound diffracting and interfering.

A room with no echo is called an anechoic chamber, and these rooms are used for experimenting with sound waves.

Sound spreads out when passing through small openings, around obstacles and through doorways. However, the effects are often owing to multiple reflections rather than diffraction.

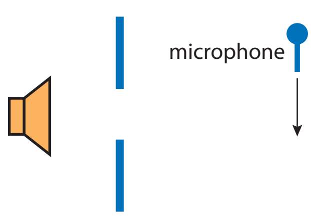


Figure 4.59 The microphone picks up sound owing to diffraction.

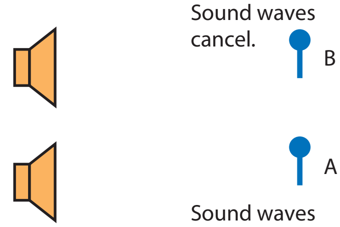


Figure 4.60 Owing to interference, the sound is loud at A but quiet at B.

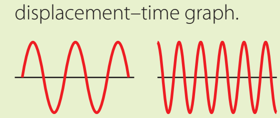
Sound has the properties of a wave, so that means we can use our wave theory to model sound. Sound is a propagation of disturbance in air pressure. Sound is an example of a longitudinal wave. The speed of sound in air is $330\ \mathrm{m\,s^{-1}}$.

- Frequency and amplitude of sound

Different frequency sound waves have different pitch (that is, a high note has a high frequency). The loudness of a sound is related to the amplitude of the wave.

Sound

Sound is created when the pressure of air is is varied. This change in pressure spreads out as a longitudinal wave. When a sound wave meets a microphone, it causes it to vibrate. The microphone then changes this vibration to an electrical signal that can be used to plot a graph. The graph that we see is a displacement–time graph.



4.6 Standing (stationary) waves

Assessment statements
- 11.1.1 Describe the nature of standing (stationary) waves.
- 11.1.2 Explain the formation of one-dimensional standing waves.
- 11.1.3 Discuss the modes of vibration of strings and air in open and in closed pipes.
- 11.1.4 Compare standing waves and travelling waves.
- 11.1.5 Solve problems involving standing waves.

To create standing waves in a string using the PhET simulation 'Waves in strings', visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 4.8.

If you take the end of a rope and flick it, you can create a wave pulse that travels along the rope. When the wave gets to the end of the rope it reflects and a pulse can be seen travelling back to your hand. If a continuous wave is sent along the rope, the original wave and reflected wave superpose to produce a wave where the peaks simply move up and down but do not progress along the rope. This is called a standing wave. Standing waves occur whenever two identical waves travelling in opposite directions superpose. Many musical instruments make use of standing waves in strings or columns of air.

Formation of standing waves

Consider two identical waves travelling along a string in opposite directions as shown in Figure 4.61.

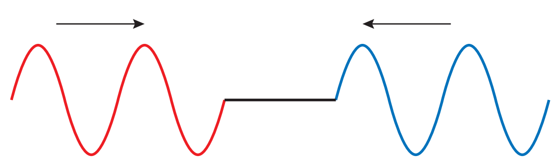


Figure 4. 61 Two waves travel towards each other.

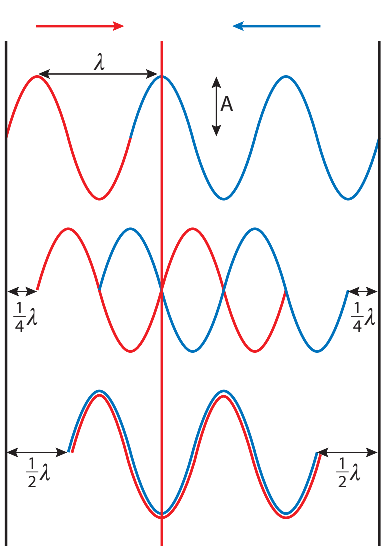


Figure 4.62 Reflected waves and original wave cross each other.

As the waves cross each other we see that they are alternately in phase (adding up) and out of phase (cancelling out). Figure 4.62 shows the two waves as they move past each other. At first they are in phase but after each has travelled $\frac{1}{4}\lambda$ they are out of phase, and after a further $\frac{1}{4}\lambda$, they are in phase again. Note the position marked with a red line; the waves add to form a peak, then add to form a trough. The resulting wave has peaks that go up and down but do not progress. The maximum and minimum displacements of the string are shown in Figure 4.63. Notice that the amplitude is 2 × the amplitude of the component waves. There are some places on the wave that don't move at all; these are called *nodes* and the parts with maximum amplitude are called *antinodes*.

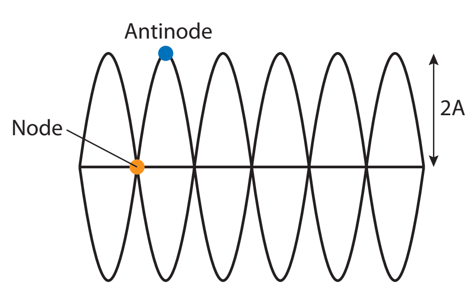


Figure 4.63

Differences between progressive waves and standing waves

The most obvious difference between a wave that travels along the rope and the standing wave is that the wave profile of a standing wave doesn't progress. Secondly, all points in between two nodes on a standing wave are in phase (think of the skipping rope) whereas points on a progressive wave that are closer than one wavelength are all out of phase (think of a Mexican wave). The third difference is related to the amplitude. All points on a progressive wave have the same amplitude, but on a standing wave some points have zero amplitude (nodes) and some points have large amplitude (antinodes).

The two ends of the skipping rope are the nodes.

Stringed instruments

Many musical instruments (guitar, violin, piano) use stretched strings to produce sound waves. When the string is plucked, a wave travels along the string to one of the fixed ends. When it gets to the end, it reflects back along the string superposing with the original wave; the result is a standing wave. The important thing to realise about the standing wave in a stretched string is that since the ends cannot move, they must become nodes, so only standing waves with nodes at the ends can be produced. Figure 4.64 shows some of the possible standing waves that can be formed in a string of length L.

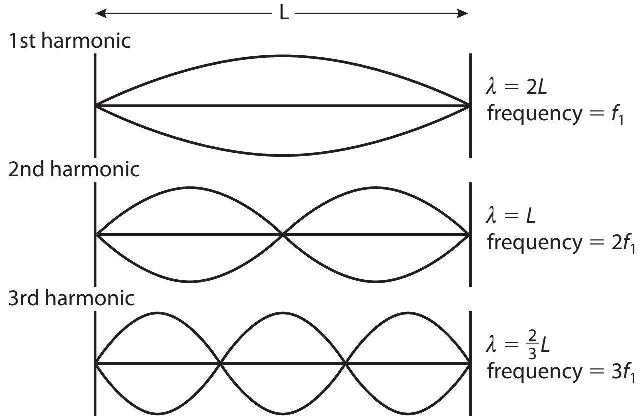


Figure 4. 64 Standing waves in a string.

As shown in the diagram, each of the possible waves is called a harmonic. The first harmonic (sometimes called the fundamental) is the wave with the lowest possible frequency. To calculate the frequency, we can use the formula $f = \frac{v}{\lambda}$ so for the first harmonic:

$$f_1 = \frac{v}{2L}$$

For the second harmonic:

$$f_2 = \frac{v}{L}$$

The wave velocity is the same for all harmonics, so we can deduce that $f_2 = 2f_1$

Wave speed

The speed of a wave in a string is given by the formula

$$v = \sqrt{\frac{T}{\mu}}$$

where T = tension and μ = mass per unit length.

This is why a thick guitar string is a lower note than a thin one and why the note gets higher when you increase the tension.

Sound produced by the guitar
Here we are focusing on the vibrating string. The string will cause the body of the guitar to vibrate which in turn causes the pressure of the air to vary. It is the pressure changes in the air that cause the sound wave that we hear.

Playing the guitar

When the guitar string is plucked, it doesn't just vibrate with one frequency but with many frequencies. However the only ones that can create standing waves are the ones with nodes at the ends (as shown in Figure 4.64). You can try this with a length of rope; get a friend to hold one end and shake the other. When you shake the end at certain specific frequencies you get a standing wave, but if the frequency isn't right, you don't. This is an example of resonance; hit the right frequency and the amplitude is big. So when the guitar string is plucked, all the possible standing waves are produced. If the signal from an electric guitar pickup is fed into a computer, the frequencies can be analysed to get a frequency spectrum. Figure 4.65 shows an example.

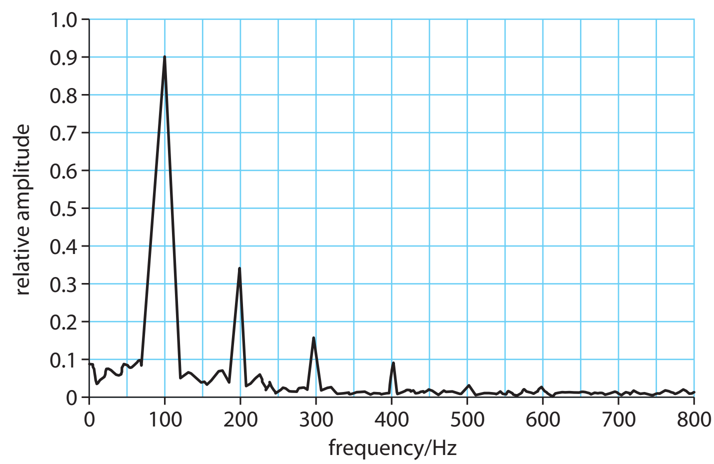


Figure 4. 65 Frequency spectrum for a string.

You can see from this graph that the string is vibrating at 100 Hz, 200 Hz, 300 Hz and so on. However, the largest amplitude note is the first harmonic (100 Hz) so this is the frequency of the note you hear.

Playing different notes

A guitar has six strings. Each string is the same length but has a different diameter and therefore different mass per unit length. The wave in the thicker strings is slower than in the thin strings so, since $f = \frac{v}{\lambda}$, the thick strings will be lower notes.

To play different notes on one string, the string can be shortened by placing a finger on one of frets on the neck of the guitar. Since $f = \frac{v}{\lambda}$ the shorter string will be a higher note.

An alternative way to play a higher note is to play a harmonic; this is done by placing a finger on the node of a harmonic (e.g. in the middle for the second harmonic) then plucking the string. Immediately after the string is plucked, the finger is removed to allow the string to vibrate freely.

Exercise

14 The mass per unit length of a guitar string is 1.2×10^{-3} kg m^{-1}. If the tension in the wire is 40 N,

(a) calculate the velocity of the wave

(b) calculate the frequency of the first harmonic if the vibrating length of the guitar string is 63.5 cm.

Standing waves in closed pipes

When a sound wave travels along a closed pipe, it will reflect off the closed end. The reflected wave and original wave then superpose to give a standing wave. A sound wave is a propagation of disturbance in air pressure. The change in air pressure causes the air to move backwards and forwards in the direction of the propagation. If the end of the pipe is closed, then the air cannot move back and forth so a node must be formed. This limits the possible standing waves to the ones shown in Figure 4.66.

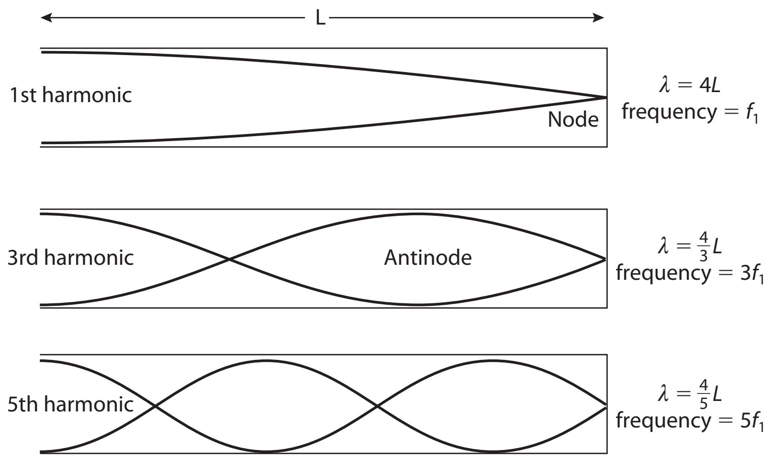


Figure 4. 66 Standing waves in a closed pipe.

These diagrams show how the displacement of the air varies along the length of the pipe, but remember, the displacement is actually along the pipe not perpendicular to it as shown. The frequency of each harmonic can be calculated using $f = \frac{v}{\lambda}$.

For the first harmonic:

$$f_1 = \frac{v}{4L}$$

For the next harmonic:

$$f_3 = \frac{v}{\frac{4}{3}L} = \frac{3v}{4L} = 3f_1 \text{ so this is the third harmonic.}$$

So when a standing wave is formed in a closed pipe, only odd harmonics are formed, resulting in the frequency spectrum shown in Figure 4.68.

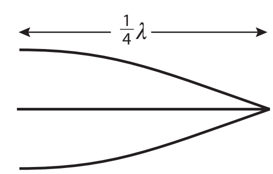


Figure 4. 67 Remember this is $\frac{1}{4}$ of a wave. It can be useful to split the harmonics into quarters when determining the frequency.

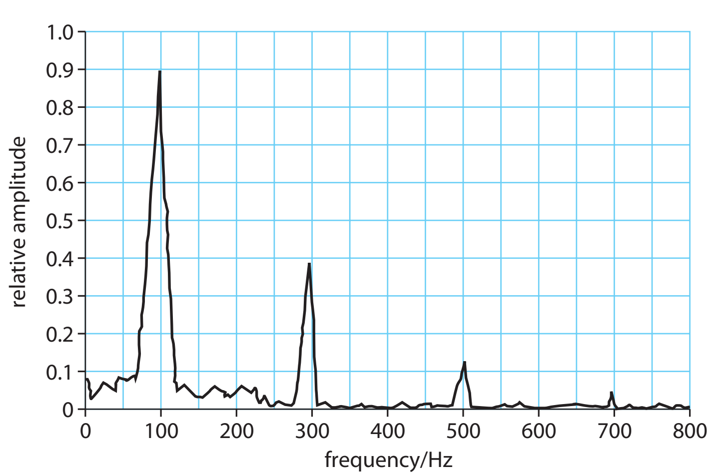


Figure 4.68 Frequency spectrum for a closed pipe.

Standing waves in open pipes

If a wave is sent along an open-ended pipe, a wave is also reflected. The resulting superposition of reflected and original waves again leads to the formation of a standing wave. This time there will be an antinode at both ends, leading to the possible waves shown in Figure 4.69.

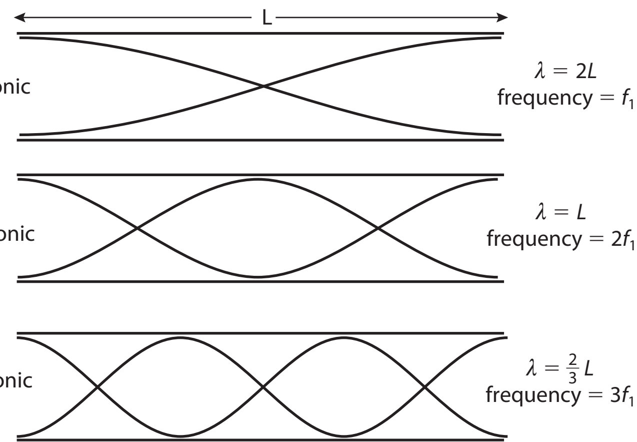


Figure 4.69 Standing waves in an open pipe.

This time all the harmonics are formed.

In a clarinet, the reed is made to vibrate.

Wind instruments

All wind instruments (e.g. flute, clarinet, trumpet and church organ) make use of the standing waves set up in pipes. The main difference between the different instruments is the way that the air is made to vibrate. In a clarinet a thin piece of wood (a reed) is made to vibrate, in a trumpet the lips vibrate and in a flute, air is made to vibrate, as it is blown over a sharp edge. Different notes are played by opening and closing holes; this has the effect of changing the length of the pipe. You can also play higher notes by blowing harder; this causes the higher harmonics to sound louder, resulting in a higher frequency note. If you have ever played the recorder, you might have had problems with this – if you blow too hard you get a high-pitched noise that doesn't sound so good.

Exercises

The speed of sound in air is 330 m s^{-1}.

15 Calculate the first harmonic produced when a standing wave is formed in a closed pipe of length 50 cm.

16 The air in a closed pipe in Figure 4.70 is made to vibrate by holding a tuning fork of frequency 256 Hz over its open end. As the length of the pipe is increased, loud notes are heard as the standing wave in the pipe resonates with the tuning fork.

(a) What is the shortest length that will cause a loud note?

(b) If the pipe is 1.5 m long, how many loud notes will you hear as the plunger is withdrawn?

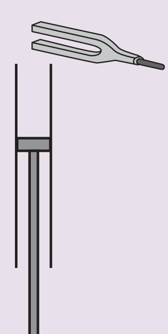


Figure 4.70

4.7 The Doppler effect

Assessment statements

11.2.1 Describe what is meant by the Doppler effect.
11.2.2 Explain the Doppler effect by reference to wavefront diagrams for moving-detector and moving-source situations.
11.2.3 Apply the Doppler effect equations for sound.
11.2.4 Solve problems on the Doppler effect for sound.
11.2.5 Solve problems on the Doppler effect for electromagnetic waves using the approximation $\Delta f = \frac{v}{c}f$
11.2.6 Outline an example in which the Doppler effect is used to measure speed.

If you have ever stood next to a busy road or even better, a racing track, you might have noticed that the cars sound different when they come towards you and when they go away. It's difficult to put this into words but the sound is something like this "eeeeeeeeeeeoowwwwwwwww". The sound on approach is a higher frequency than on retreat. This effect is called the Doppler effect, and can occur when the source of the sound is moving or when the observer is moving, or when both source and observer are moving.

When we say the source is moving we mean that it moves relative to the medium (air). The observer is at rest relative to the medium.

Moving source

The change in frequency caused when a source moves is due to the change in wavelength in front of and behind the source. This is illustrated in Figure 4.71. You can see that the waves ahead of the source have been squashed as the source 'catches up' with them. The velocity of sound is not affected by the movement of the source, so the reduction in wavelength results in an increased frequency ($v = f\lambda$).

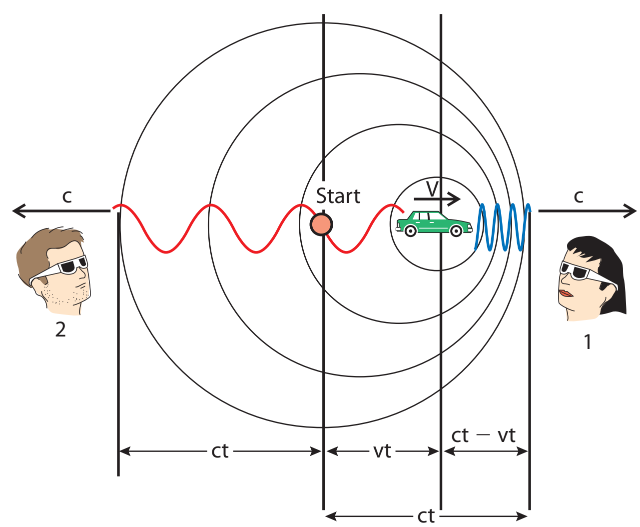


Figure 4.71 The car starts from the red spot and moves forward. The largest circle is the wavefront formed when the car began.

The car in Figure 4.71 starts at the red spot and drives forwards at speed v for t seconds producing a sound of frequency f_0. In this time t the source produces f_0t complete waves ($3\frac{1}{2}$ in this example).

Ahead of the car, these waves have been squashed into a distance $ct - vt$ so the wavelength must be

$$\lambda_1 = \frac{ct - vt}{f_0t} = \frac{c - v}{f_0}$$

The velocity of these waves is c so the observed frequency $f_1 = \frac{c}{\lambda_1}$

$$f_1 = \frac{c}{\frac{c - v}{f_0}} = \frac{cf_0}{c - v}$$

We could do a similar derivation for the waves behind the source. This time the waves are stretched out to fit into a distance $ct + vt$. The observed frequency is then

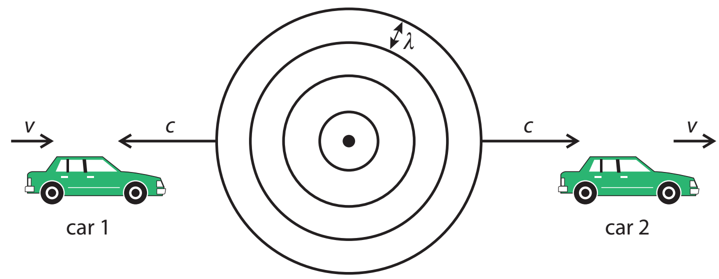
$$f_2 = \frac{cf_0}{c + v}$$

If the source moves at the speed of sound the sound in front bunches up to form a shock wave. This causes the bang (sonic boom) you hear when a plane breaks the sound barrier.

Moving observer

If the observer moves relative to a stationary source the change in frequency observed is simply due to the fact that the velocity of the sound changes. This is because the velocity of sound is relative to the air, so if you travel through the air towards a sound, the velocity of the sound will increase. Figure 4.72 illustrates the effect on frequency of a moving observer.

Figure 4. 72



The relative velocity of the sound coming towards car 1 approaching the source is $(c + v)$ so the frequency, $f_1 = \frac{c+v}{\lambda}$.

But $$\lambda = \frac{c}{f_0}$$

so $$f_1 = \frac{(c+v)f_0}{c} \quad \text{(approaching)}$$

For car 2, receding from the source

$$f_2 = \frac{(c-v)f_0}{c} \quad \text{(receding)}$$

Worked example

A car travelling at 30 ms^{-1} emits a sound of frequency 500 Hz. Calculate the frequency of the sound measured by an observer in front of the car.

Solution

This is an example where the source is moving relative to the medium and the observer is stationary relative to the medium. To calculate the observed frequency, we use the equation

$$f_1 = \frac{cf_0}{(c-v)}$$

where
$$f_0 = 500\ \text{Hz}$$
$$c = 330\ \text{ms}^{-1}$$
$$v = 30\ \text{ms}^{-1}$$

So $$f_1 = \frac{330 \times 500}{330 - 30} = 550\ \text{Hz}$$

Exercises

17 A person jumps off a high bridge attached to a long elastic rope (a bungee jump). As they begin to fall they start to scream at a frequency of 1000 Hz. They reach a terminal velocity of 40 m s^{-1} on the way down and 30 m s^{-1} on the way back up.

(a) Describe what you would hear if you were standing on the bridge.
(b) Calculate the maximum frequency you would hear from the bridge.
(c) Calculate the minimum frequency you would hear from the bridge.
(d) What would you hear if you were standing directly below the bungee jumper?

18 The highest frequency you can hear is 20 000 Hz. If a plane making a sound of frequency 500 Hz went fast enough, you would not be able to hear it. How fast would the plane have to go?

19 Calculate the frequency of sound you would hear as you drove at 20 m s^{-1} towards a sound source emitting a sound of frequency 300 Hz.

The Doppler effect and EM radiation

The Doppler effect also applies to electromagnetic radiation (radio, microwaves and light). The derivation of the formula is rather more complicated since the velocity of light is not changed by the relative movement of the observer. However if the relative velocities are much smaller than the speed of light we can use the following approximation:

$$\Delta f = \frac{v}{c} f_0$$

where Δf = the change in frequency
v = the relative speed of the source and observer
c = the speed of light in a vacuum
f_0 = the original frequency

Hint: Unlike the Doppler effect for sound, it doesn't matter whether it is the source or observer that moves.

Red shift

If a source of light is moving away from an observer, the light received by the observer will have a longer wavelength than when it was emitted. If we look at the spectrum as shown in Figure 4.73, we see that the red end of the spectrum is long wavelength and the blue end is short wavelength. The change in wavelength will therefore cause the light to shift towards the red end of the spectrum – this is called 'red shift'. This effect is very useful to astronomers as they can use it to calculate how fast stars are moving away from us. To do this calculation they have to know the original wavelength of the light. They can work this out because the spectrum of each element is like a fingerprint, so it is possible from the fingerprint to work out what the stars are made of and therefore which frequencies of light they emit.

Red shift is due to the expansion of the universe rather than the relative motion of source and observer. However, the effect is the same.

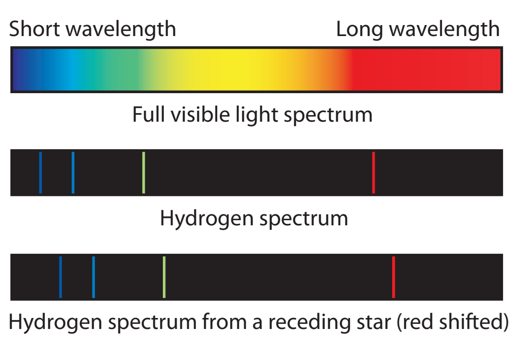


Figure 4.73 Comparing the spectrum of hydrogen from a stationary source and a star.

Speed traps

Police speed traps use the Doppler effect to measure the speed of passing cars. When they aim the device at the car, a beam of EM radiation (microwaves or IR) is reflected off the car. The reflected beam undergoes a double Doppler shift; first the car is approaching the beam so there will be a shift due to the moving observer and secondly the reflected beam is emitted from the moving source. When the device receives the higher frequency reflected beam, the speed of the car can be calculated from the change in frequency.

Speed cameras use the Doppler effect.

Worked example

A police speed trap uses a beam with a wavelength of 904 nm. What frequency will be received by a car travelling towards the trap at 150 km h^{-1}?

Solution

First convert the car's speed to m s^{-1}

150 km h^{-1} = 42 m s^{-1}

The change in frequency is then found using the formula

$$\Delta f = \frac{v}{c} f_0$$

The frequency of the signal, $f_0 = \frac{c}{\lambda} = \frac{3 \times 10^8}{904 \times 10^{-9}}$

$$f_0 = 3.3 \times 10^{14}\ \text{Hz}$$

So $\Delta f = \frac{42 \times 3.3 \times 10^{14}}{3 \times 10^8}$

$= 4.6 \times 10^7$ Hz

Since the car is travelling towards the source, this frequency will be added on to the original 3.3×10^{14} Hz.

The Doppler shift equation can be also be written in terms of the wavelength

$$\Delta\lambda = \frac{v}{c}\lambda_0$$

Exercises

20 A star emits light of wavelength 650 nm. If the light received at the Earth from this star has a wavelength of 690 nm, how fast is the star moving away from the Earth?

21 If the car in example 1 reflects the received signal, what will the total Doppler shift in the signal be when it arrives back to the policeman?

22 An atom of hydrogen travelling towards the Earth at 2×10^6 m s^{-1} emits light of wavelength 658 nm. What is the change in wavelength experienced by an observer on the Earth?

4.8 Diffraction at a single slit

Assessment statements

11.3.1 Sketch the variation with angle of diffraction of the relative intensity of light diffracted at a single slit.

11.3.2 Derive the formula $\theta = \frac{\lambda}{b}$ for the position of the first minimum of the diffraction pattern produced at a single slit.

11.3.3 Solve problems involving single-slit diffraction.

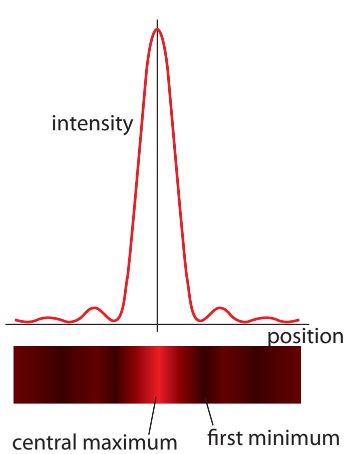


Figure 4.74

When light passes through a narrow slit, it spreads out due to diffraction. The spreading out is not uniform but forms a pattern, as shown in Figure 4.74. To derive an equation for this pattern we must first understand how light propagates.

Huygens construction

In 1690 the Dutch physicist Christian Huygens (developing an idea of Robert Hooke's) proposed a model for the propagation of light. He suggested that light propagated as if the wavefront were made of an infinite number of small wavelet sources, each wavelet source producing a small spherical wavelet in the forward direction. The new wavefront was the sum of all these wavelets. Figure 4.75 shows how plane and circular wavefronts propagate according to this model. Using this model we can find out how a plane wavefront propagates through a narrow slit. If the slit is very narrow the solution is easy, since only one wavelet would go through, resulting in a circular wavefront. However, if the slit is wider we have to find the sum of all the wavelets.

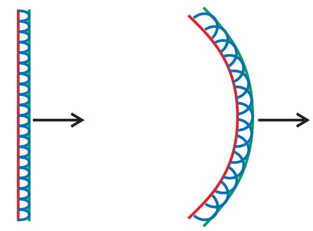


Figure 4. 75 Huygens construction used to find the new position of plane and circular wavefronts.

Applying Huygens construction to a narrow slit

When a wavefront passes through a narrow slit, it will propagate as if there were a large number of wavelet sources across the slit as in Figure 4.76.

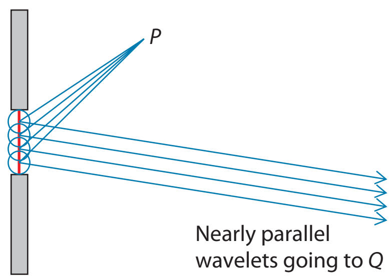


Figure 4.76 Wavelets add to give resultant intensity.

The resultant intensity at some point P in front of the slit is found by summing all the wavelets. This is not a simple matter, since each wavelet has travelled a different distance so they will be out of phase when they arrive at P. To simplify the problem, we will consider a point Q a long way from the slits. Light travelling through the slit arriving at point Q is almost parallel, and if we say that it is parallel, then the geometry of the problem becomes much simpler.

W To see how waves propagate, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 4.9.

The central maximum

The central maximum occurs directly ahead of the slit. If we take a point a long way from the slit, all wavelets will be parallel and will have travelled the same distance, as shown in Figure 4.77. If all the wavelets have travelled the same distance they will be in phase, so will add up to give a region of high intensity (bright).

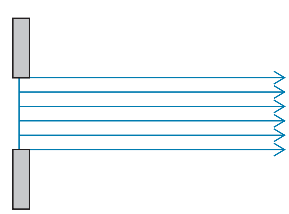


Figure 4.77 Wavelets travelling to the central maximum.

The first minimum

If we now consider wavelets travelling towards the first minimum, as in Figure 4.78, they are travelling at an angle so will not all travel the same distance. The wavelet at the top will travel further than the wavelet at the bottom. When these wavelets add together they cancel each other out to form a region of low intensity (dark).

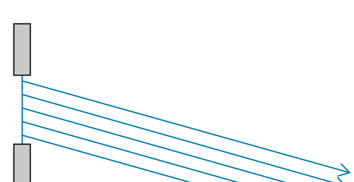


Figure 4.78 Wavelets travelling to the first minimum.

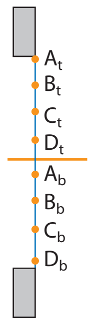


Figure 4.79 Wavelets in the top half cancel with wavelets in the bottom.

We can calculate the angle at which the first minimum is formed by splitting the slit into two halves, top and bottom, as in Figure 4.79. If all the wavelets from the top half cancel out all the wavelets from the bottom, the result will be a dark region. So if we have 8 wavelet sources, 4 in the top half (A_t B_t C_t D_t) and 4 in the bottom (A_b,B_b,C_b,D_b) and if A_t cancels with A_b and B_t cancels with B_b etc., then all the wavelets will cancel with each other. For each pair to cancel, the path difference must be $\frac{1}{2}\lambda$. Figure 4.80 shows the situation for the top wavelet and the one in the middle.

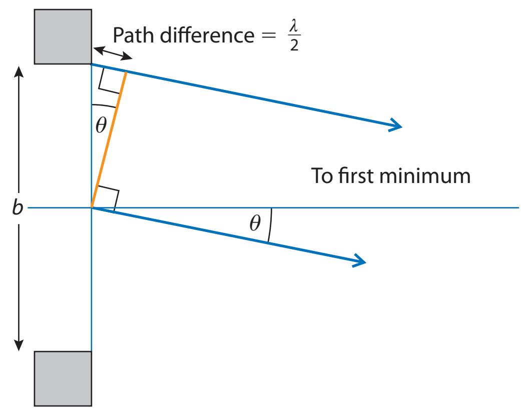


Figure 4.80 Geometric construction for the first minimum.

The orange line cuts across the two wavelets at 90° showing that the top one is longer than the bottom. If the path difference shown is $\frac{\lambda}{2}$ then these wavelets will cancel and so will all the others. If the first minimum occurs at an angle θ as shown, then this will also be the angle of the triangle made by the orange line. We can therefore write:

$$\sin\theta = \frac{\frac{\lambda}{2}}{\frac{b}{2}} = \frac{\lambda}{b}$$

But the angles are very small, so if θ is measured in radians, $\sin\theta = \theta$,

So: $$\theta = \frac{\lambda}{b}$$

Knowing the position of the first minimum tells us how spread out the diffraction pattern is. From the equation we can see that if b is small then θ is big, so the pattern is spread out as shown in Figure 4.81.

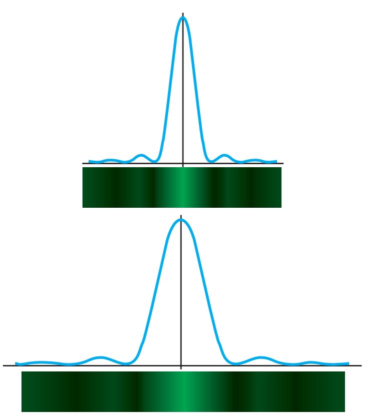
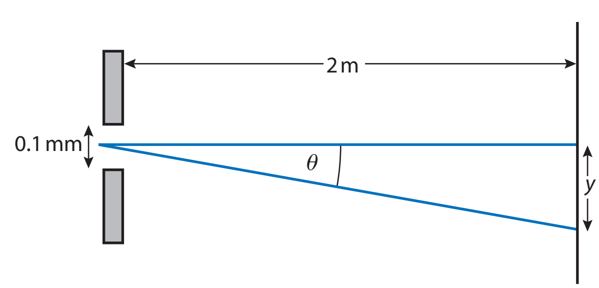


Figure 4.81 Comparing a wide slit (top) with a narrow slit (bottom).

Example

A diffraction pattern is formed on a wall 2 m from a 0.1 mm slit illuminated with light of wavelength 600 nm. How wide will the central maximum be?

First draw a diagram showing the relative positions of the slit and screen.



The angle θ can be calculated from $\theta = \frac{\lambda}{b}$

$$= \frac{600 \times 10^{-9}}{0.1 \times 10^{-3}}$$

$$= 0.006 \text{ rads}$$

Since this angle is small we can say that $\theta = \frac{y}{2}$ m so $y = 0.006 \times 2$ m

$$= 0.012 \text{ m}$$

$$= 1.2 \text{ cm}$$

This is half the width of the maximum, so width $= 2.4$ cm

Exercises

23 Light of wavelength 550 nm is passed through a slit of size 0.05 mm. Calculate the width of the central maximum formed on a screen that is 5 m away.

24 Calculate the size of the slit that would cause light of wavelength 550 nm to diffract forming a diffraction pattern with a central maximum 5 cm wide on a screen 4 m from the slit.

4.9 Resolution

Assessment statements

11.4.1 Sketch the variation with angle of diffraction of the relative intensity of light emitted by two point sources that has been diffracted at a single slit.

11.4.2 State the Rayleigh criterion for an image of two sources to be just resolved.

11.4.3 Describe the significance of resolution in the development of devices such as CDs and DVDs, the electron microscope and radio telescopes.

11.4.4 Solve problems involving resolution.

In the previous section we discovered that when light passes through a narrow slit it spreads out. This can cause problems for anyone using a device, such as a camera, telescope or even your eye, where light passes through a narrow opening. Try looking in the square of Figure 4.82; you should be able to see that there are two dots; we say you can 'resolve' the two dots. If you move away from the page there will be a point where you can no longer resolve the two dots – they now look like one dot. The reason for this is when the light passes into your eye it diffracts; this causes the dots to appear as spots on your retina. As you move away from the page the image of the dots gets closer and closer until the two spots on your retina overlap, and the dots now look like one.

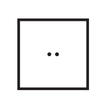
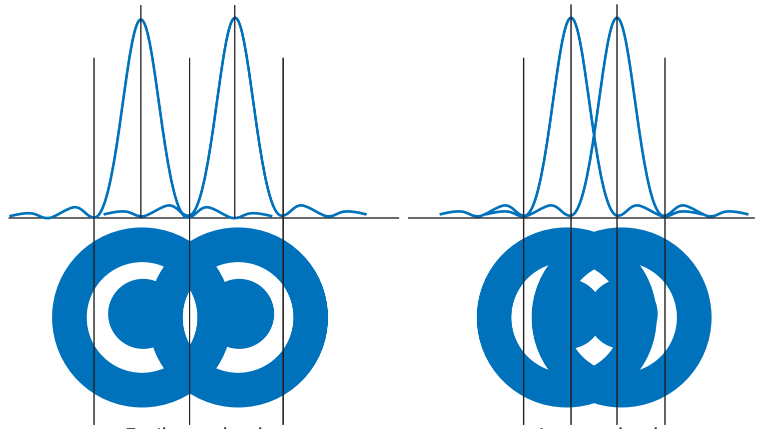


Figure 4.82

The Rayleigh criterion

The Rayleigh criterion puts a limit on the resolvability of two points, based on the diffraction of light. It states that two points will just be resolved if the central maximum of the diffraction pattern formed of one point coincides with the first minimum of the other. Figure 4.83 shows this. Note that the diffraction patterns are not made of lines like those formed by slits, but are circular. This is because the aperture at the front of most optical instruments and the eye is circular, and circular apertures form circular diffraction patterns.

Figure 4. 83 If the diffraction patterns overlap more than this, the two spots look like one.



From Figure 4.83 we can see that if the distance between the central maxima is less than half the width of the maxima the points will not be resolved. The width of the central maxima is defined by the position of the first minimum, which for a slit is given by the equation $\theta = \frac{\lambda}{b}$. But the central maximum for a circular aperture is 1.22 times wider than a slit, so the angular position of the first minimum is given by $\theta = \frac{1.22\lambda}{b}$ where $b =$ diameter. Using this equation we can calculate whether two spots will be resolved or not, as long as we know the wavelength of light and the size of the aperture.

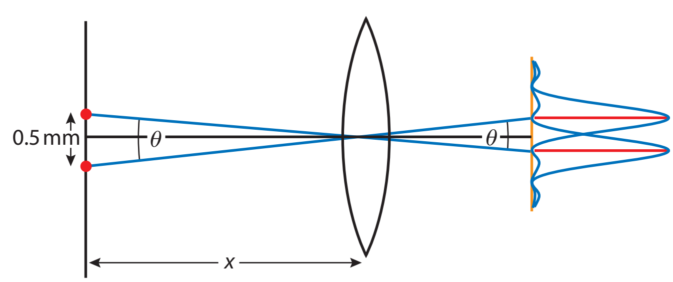
Diffraction is not the only factor that affects the resolution of objects by optical instruments. The quality of the lens and the number of pixels in a digital camera are often the limiting factors.

Worked example

Two small red spots 0.5 mm apart reflect light of wavelength 650 nm into a camera that has an aperture of 5 mm. What is the maximum distance between the camera and the paper if the spots are to be resolved in the photograph?

Solution

First draw a diagram:



The diagram shows two rays of light passing through the centre of the lens landing on the film. Notice these rays do not change direction since they pass through the centre of the lens. Light from each point forms a diffraction pattern on the film. If the separation of the patterns equals half the width of the central maximum of one, they will just be resolved.

$$\text{The angle } \theta \text{ on the right} = \frac{1.22\lambda}{b}$$

$$= \frac{1.22 \times 650 \times 10^{-9}}{5 \times 10^{-3}}$$

$$= 1.59 \times 10^{-4}\ \text{rad}$$

The angle θ on the left is the same and since it is small

$$\theta = \frac{0.5 \times 10^{-3}}{x}$$

so $$x = \frac{0.5 \times 10^{-3}}{1.59 \times 10^{-4}}$$

$$= 3.1 \text{ m}$$

Exercises

25 A telescope with aperture diameter 4 cm is used to view a collection of stars that are 4×10^{12} m away. If the stars give out light of wavelength 570 nm, what is the distance between the closest stars that can be resolved?

26 To read the headline in a newspaper you need to resolve points that are about 1 cm apart. If a spy camera is to be put in orbit 200 km above the Earth, how big must its aperture be if it is to take readable pictures of newspaper headlines? (Assume wavelength = 600 nm.)

27 The pixels on a computer screen are about 0.01 cm apart. If the aperture of your eye has diameter 5 mm should you be able to resolve the pixels at a distance of 1 m? (Assume wavelength = 600 nm.) Explain your answer.

Improving resolution

Whenever electromagnetic radiation passes through an aperture, it is diffracted. This causes problems if you need to resolve the parts of the radiation that come from different places. We have seen how this affects optical instruments and how we can improve the situation by increasing the size of the aperture. An alternative way of increasing resolution is to use radiation of a different wavelength.

How more information is stored on a DVD than a CD

A CD is a shiny disc covered in small bumps that are called 'pits'. These pits are about 5×10^{-7} m wide. As the CD rotates in the CD reader, a laser of wavelength 780 nm is reflected off the shiny surface into a detector. If a pit goes past, the detector senses a change and sends a signal to the computer. The computer decodes this signal converting it into music, pictures or something else. To get more information onto the disc, you would need to make the pits smaller, but if you do that, they diffract the light so much that the detector could not resolve the difference between the pits and the gaps. To solve this problem, a laser of shorter wavelength (640 nm) is used.

The electron microscope

A microscope is used to give a big image of a small object, but there is no point in making a big image if you can't see the detail. This is the problem with microscopes that use light – their resolving power is limited by the aperture size and wavelength. The closest points that can be resolved with an optical microscope are about 200 nm apart. To increase the resolving power you could use shorter wavelengths, but then you can't see the image since the light would not be visible. An electron microscope uses high-energy electrons that have wavelengths as small as 0.02 nm, enabling them to resolve points as close as 0.1 nm (that's about the size of an atom). You can't see the image directly, but the electrons are detected by sensors and the image produced electronically.

Using an electron microscope, it is possible to produce an image of a single atom – but is seeing with electrons really seeing?

The radio telescope

A radio telescope detects the radio waves emitted by celestial bodies and converts them into an image. These radio waves have wavelengths of about 20 cm, so to have good resolution, the telescopes must be very big. The Lovell radio telescope at Jodrell Bank has a diameter of 76 m.

Radio telescopes need to be huge.

4.10 Polarization

Assessment statements

11.5.1 Describe what is meant by polarized light.
11.5.2 Describe polarization by reflection.
11.5.3 State and apply Brewster's law.
11.5.4 Explain the terms polarizer and analyser.
11.5.5 Calculate the intensity of a transmitted beam of polarized light using Malus' law.

As mentioned earlier, when a transverse wave is polarized, the disturbance is only in one plane. This is best demonstrated with a wave in a string. You can flick a string in any direction, sending a wave along it horizontally, vertically or anything in between. However, if the motion of the string is restricted by passing it through a vertical slit, the wave can only be vertical. It is then said to be polarized, the plane of polarization in this case being vertical. If a second vertical slit is placed on the string then the wave can pass unhindered. However, if the second slit is horizontal, the wave is stopped. If the second slit is rotated then the wave will come and go, as the slits line up every half revolution.

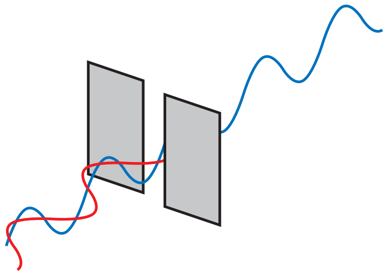


Figure 4.84 A wave in a string is polarized by a vertical slit.

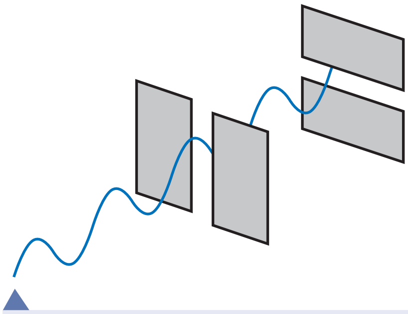


Figure 4.85 The polarized wave passes through a vertical slit but not a horizontal slit.

Athough the polarization of light is nothing to do with slits, light can still be polarized. This is achieved by passing light through a piece of special plastic called Polaroid (it's the stuff some sunglasses are made of). We can't see that the light becomes polarized, but if we pass the polarized light through a second piece of polaroid and rotate it, the light goes away every half a revolution, just like the wave in the string does. The effect is the same so we use the same model.

Even though you can't see that the light is polarized, we say it is because it does the same thing as the wave in a string. We often use models of things we can see to help us understand things that we can't.

Brewster's angle

Polaroid is used to make the lenses of sunglasses. Polarizing material can be used in this way because when the light is polarized, only the components of light in the direction of polarization are let through. This reduces the intensity of the transmitted light by 50%. Fishermen often wear polaroid sunglasses, because they reduce the glare caused by light reflected off the water, enabling them to see the fish. This works because when light is reflected, it becomes polarized in the plane of the surface. If the fisherman's sunglasses have the polarizers perpendicular to this plane then the reflected light can't pass (if the fisherman tilts his head to the side he won't be able to see the fish). The amount of polarization upon reflection on a surface such as water depends upon the angle of incidence of the light. In 1812 Sir David Brewster discovered an interesting fact:

When the angle between the reflected ray and the refracted ray is 90°, polarization is 100%. Figure 4.86 shows the angles involved.

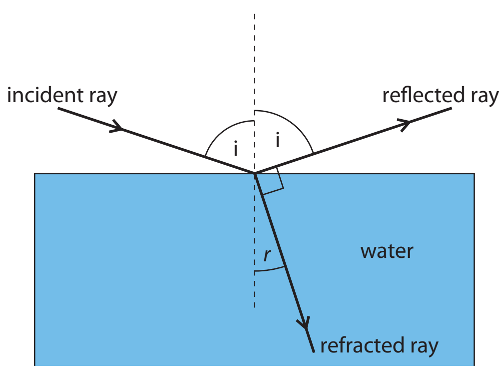


Figure 4.86 100% polarization when the rays make a 90° angle like this.

Intensity

The intensity of light is the power that is incident on a unit area. The unit of intensity is therefore $\mathrm{W\,m^{-2}}$. The greater the intensity the brighter the light.

Intensity is proportional to amplitude^2.

If we look at the angles on the right side of the normal we see that

$i + r + 90° = 180°$ so $r = 90° - i$

Now we know that Snell's law states that $\dfrac{\sin i}{\sin r} = n$ where n is the refractive index of water (since the first medium is air). Substituting for r gives

$\dfrac{\sin i}{\sin(90° - i)}$ but $\sin(90° - i) = \cos i$

so $n = \dfrac{\sin i}{\cos i} = \tan i$

This angle of incidence is called Brewster's angle φ

$$\varphi = \tan^{-1} n$$

Exercises

28 Calculate Brewster's angle for water (refractive index 1.5).

29 Explain why a fisherman fishing in the middle of the day in the summer finds his polaroid sunglasses are not very effective.

Malus's law

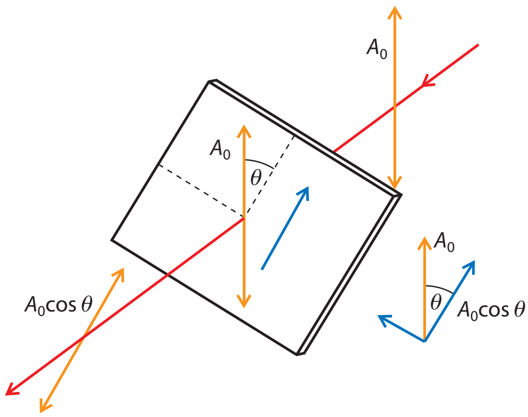


Figure 4.87 Only the component in the plane of polarization passes through.

When unpolarized light passes through a polarizer, its intensity is reduced by 50%. When this polarized light passes through a second polarizer (called an analyser), the reduction in intensity depends on the angle between the polarization planes of the two polarizers. If the two polarizers are parallel, then all the light is transmitted. If they are perpendicular, none is transmitted. If the angle between the polarizers is θ (between 0° and 90°) then some light passes. This is because the analyser will allow through the component of the polarized light that has the same plane of polarization as the analyser.

When polarized light passes through an analyser, only the component of amplitude that is in the direction of the polarizing plane of the analyser will pass through. From Figure 4.87, you can see that this component $= A_0 \cos\theta$

The intensity of light (I) is proportional to A^2 so if the original intensity of the polarized light was I_0, the intensity passing through the analyser will be given by the equation

$$I = I_0 \cos^2\theta$$

This is called Malus's law.

Exercise

30 Unpolarized light of intensity I_0 is incident on a polarizer. The transmitted light is then incident on a second polarizer. The axis of the second polarizer makes an angle of 60° to the axis of the first polarizer. Calculate the fraction of intensity transmitted.

4.11 Uses of polarization

Assessment statements

11.5.6 Describe what is meant by an optically active substance.
11.5.7 Describe the use of polarization in the determination of the concentration of certain solutions.
11.5.8 Outline qualitatively how polarization may be used in stress analysis.
11.5.9 Outline qualitatively the action of liquid crystal displays (LCDs).
11.5.10 Solve problems involving the polarization of light.

Optically active substances

Polaroid only allows through one plane of polarized light, but some substances (such as sugar solution) can rotate the plane of polarization. These are called 'optically active' substances. You can demonstrate this by setting up two crossed polarizers (polarizers with perpendicular polarizing planes). No light will pass through the crossed polarizers, but if you put some sugar solution between them it rotates the plane of polarization, so a component of the light is now able to pass.

If you rotate the second polarizer (the analyser) you could find out the new plane of polarization – it would be the angle that allows most light through; different wavelengths are rotated by different amounts so as the analyser is rotated, the sugar would appear different colours.

Measuring concentration
The angle of rotation is related to the concentration of sugar solution so concentration can be measured by using the set-up as in Figure 4.88 and rotating the analyser until the light passes with maximum brightness. The angle of the analyser gives the angle of rotation from which the concentration can be calculated.

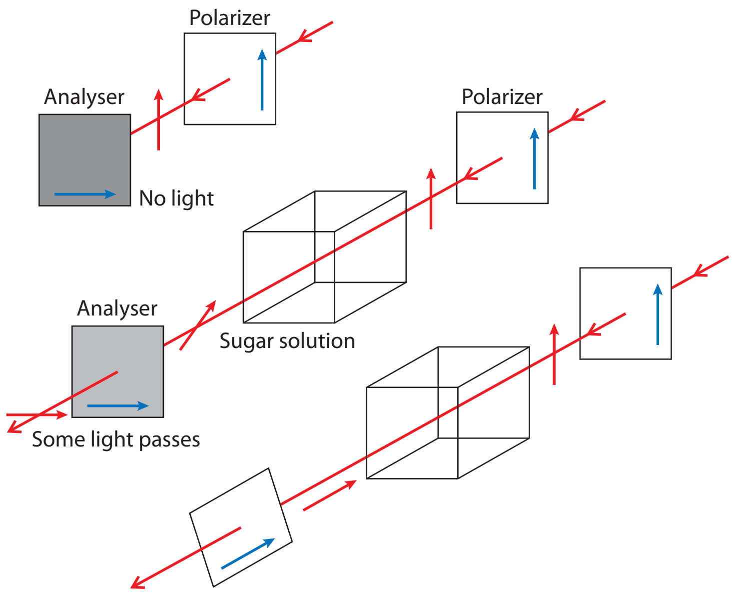
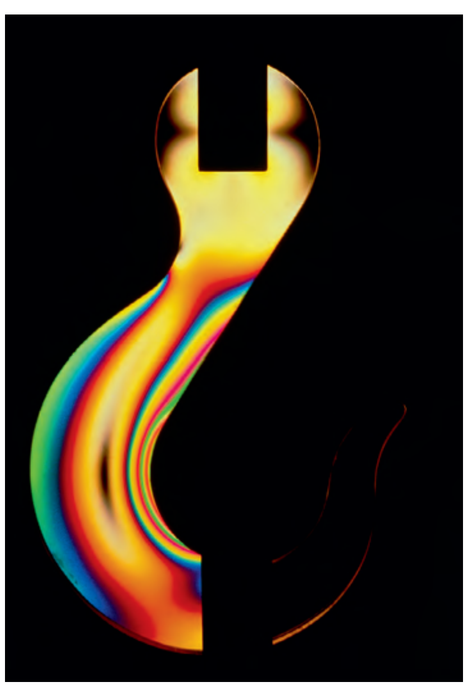


Figure 4.88 Plane of polarization rotated by sugar solution.

Some plastics also become optically active when subjected to force (put under stress). The angle of rotation for different colours is dependent on the amount of force that the plastic is subjected to. Here a copy of a crane hook has been made out of plastic. When viewed in polarized light, coloured patterns are revealed that show the areas that are subjected to the greatest stress. By studying these pictures, engineers can change the design of the hook to spread the load evenly, eliminating stress points.



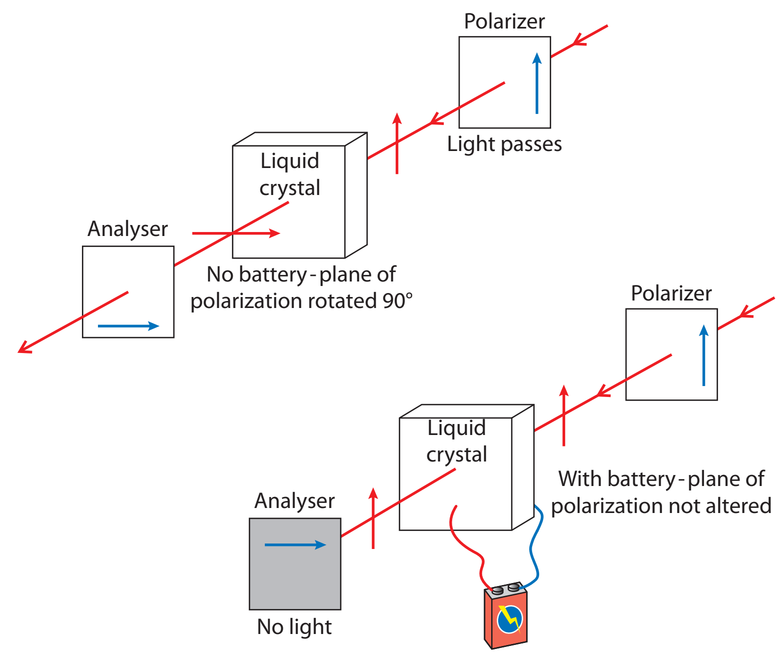
When viewed through crossed polarizers, different areas have different colours as shown in the photo.

Liquid crystal displays (LCD)

Calculators, watches, computer screens and televisions have displays that are made up of thousands of small dots called pixels. In an LCD, each pixel is made of a tiny liquid crystal. Liquid crystals have a very useful property; normally they rotate the plane of polarization through 90°, but when a battery is connected across them, they don't. So if a liquid crystal is placed between two crossed polarizers as shown in Figure 4.89, the crystal goes dark when the battery is connected.

Try looking at your mobile phone, calculator or computer screen whilst wearing Polaroid sunglasses. You may have to rotate your head to see the effect.

Figure 4.89 How an LCD works.



Front lit displays
Figure 4.89 shows how a back lit display works; these are used on computers and TVs. A lot of calculators and watches do not have back lighting but use the light from the room to illuminate the display. In this case a mirror is placed behind the analyser. The light reflects back if there is no battery connected, but does not reflect when connected.

A seven segment LCD – each segment is a separate crystal.

To make a display, thousands of these small crystals are connected in a grid. A picture is made by applying a potential difference (connecting a battery) to selected pixels. How this works is very easy to see in the old calculator displays that used just seven segments to form the numbers from 1 to 9.

Exercise

31 Light of intensity 10 W m^{-2} passes through two crossed polaroids.

(a) What is the intensity of light after the first polaroid?

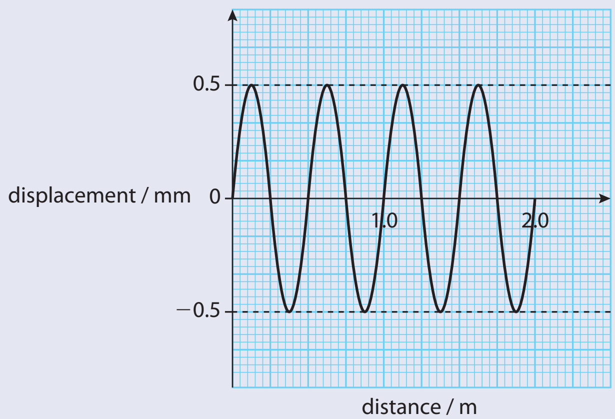
(b) What is the intensity after the second polaroid?

A sugar solution is placed between the two polaroids and rotates the plane of polarization through 30°.

(c) What is the intensity of the light that now emerges from the second polaroid?

Practice questions

1 This question is about sound waves.
A sound wave of frequency 660 Hz passes through air. The variation of particle displacement with distance along the wave at one instant of time is shown below.



(a) State whether this wave is an example of a longitudinal or a transverse wave. (1)

(b) Using data from the above graph, deduce for this sound wave:

(i) the wavelength. (1)

(ii) the amplitude. (1)

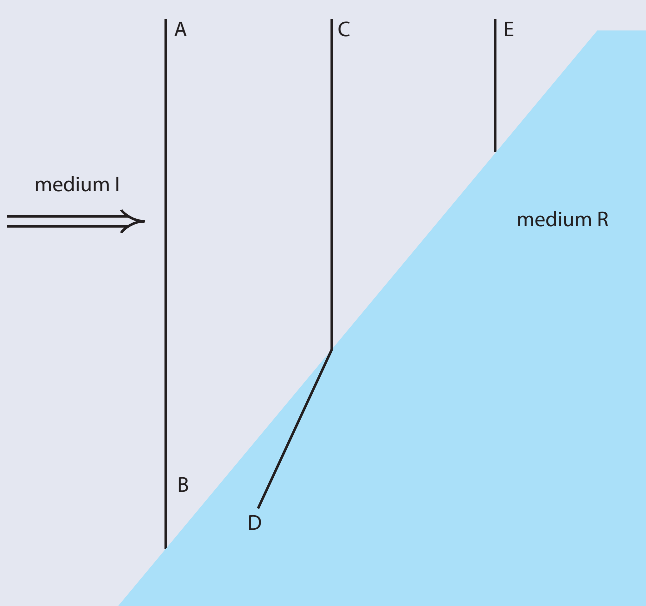
(iii) the speed. (2)

2 This question is about waves and wave properties.

(a) By making reference to waves, distinguish between a *ray* and a *wavefront*. (3)

The following diagram shows three wavefronts incident on a boundary between medium I and medium R. Wavefront CD is shown crossing the boundary. Wavefront EF is incomplete.

(b)

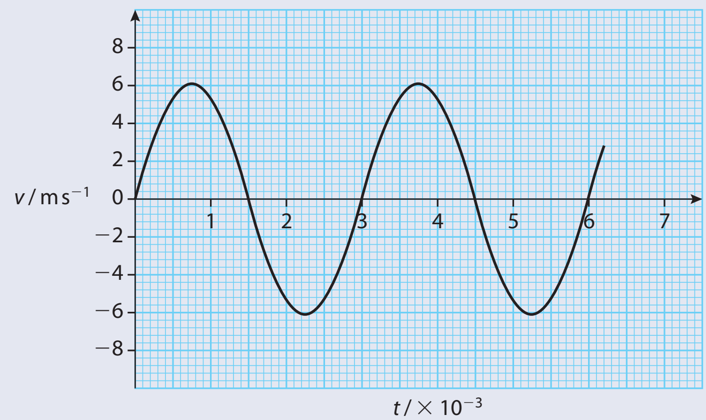


(i) On the diagram above, draw a line to complete the wavefront EF. (1)

(ii) Explain in which medium, I or R, the wave has the higher speed. (3)

The graph below shows the variation with time t of the velocity v of one particle of the medium through which the wave is travelling.

(c)



(i) Explain how it can be deduced from the graph that the particle is oscillating. (2)

(ii) Determine the frequency of oscillation of the particle. (2)

(iii) Mark on the graph with the letter M one time at which the particle is at maximum displacement. (1)

(iv) Estimate the area between the curve and the x-axis from the time $t = 0$ to the time $t = 1.5$ ms.

(v) Suggest what the area in **c(iv)** represents. (3)

3 This question is about waves and wave motion.

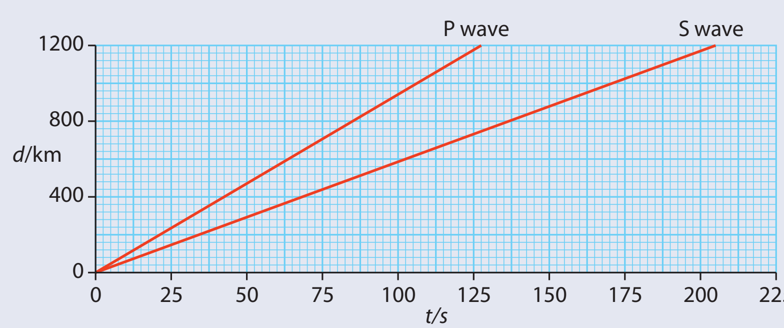
(a) **(i)** Define what is meant by the *speed of a wave*. (2)

(ii) Light is emitted from a candle flame. Explain why, in this situation, it is correct to refer to the 'speed of the emitted light', rather than its velocity. (2)

(b) **(i)** Define, by reference to wave motion, what is meant by *displacement*. (2)

(ii) By reference to displacement, describe the difference between a longitudinal wave and a transverse wave. (3)

The centre of an earthquake produces both longitudinal waves (P waves) and transverse waves (S waves). The graph below shows the variation with time t of the distance d moved by the two types of wave.

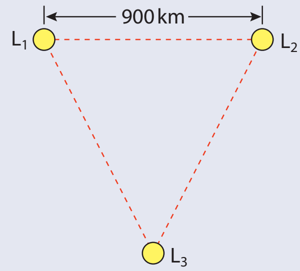


(c) Use the graph to determine the speed of

(i) the P waves. (1)

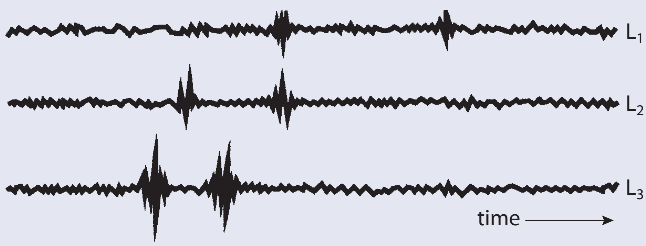
(ii) the S waves. (1)

(d) The waves from an earthquake close to the Earth's surface are detected at three laboratories L_1, L_2 and L_3. The laboratories are at the corners of a triangle so that each is separated from the others by a distance of 900 km, as shown in the diagram below.



(2)

The records of the variation with time of the vibrations produced by the earthquake as detected at the three laboratories are shown below. All three records were started at the same time.



On each record, one pulse is made by the S wave and the other by the P wave. The separation of the two pulses is referred to as the S-P interval.

(i) On the trace produced by laboratory L_2, identify, by reference to your answers in **(c)**, the pulse due to the P wave (label the pulse P). (1)

(ii) Using evidence from the records of the earthquake, state which laboratory was closest to the site of the earthquake. (1)

(iii) State **three** separate pieces of evidence for your statement in **(d)(ii)**. (3)

(iv) The S-P intervals are 68 s, 42 s and 27 s for laboratories L_1, L_2 and L_3 respectively. Use the graph, or otherwise, to determine the distance of the earthquake from each laboratory. Explain your working.

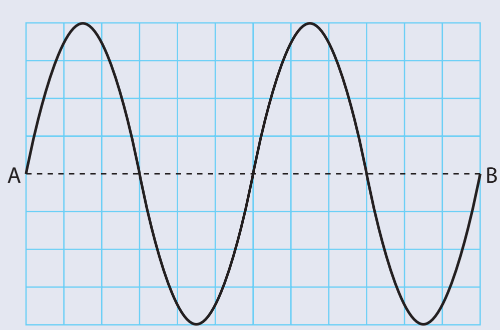
(v) Mark on the diagram a possible site of the earthquake.

4 This question is about the interference of waves.

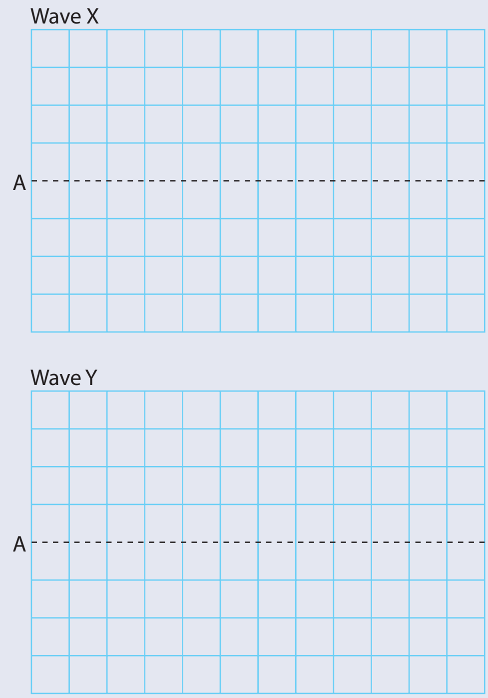
(a) State the principle of superposition. (2)

A wire is stretched between two points A and B.

A standing wave is set up in the wire. This wave can be thought of as being made up from the superposition of two waves, a wave X travelling from A to B and a wave Y travelling from B to A. At one particular instant in time, the displacement of the wire is as shown. A background grid is given for reference and the equilibrium position of the wire is shown as a dotted line.

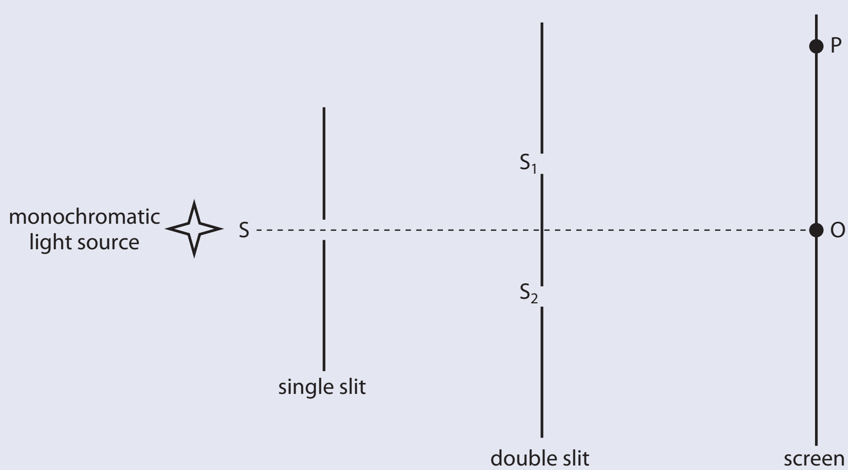


(b) On the grids below, draw the displacement of the wire due to wave X and wave Y.



(4)

The diagram below shows an arrangement (not to scale) for observing the interference pattern produced by the superposition of two light waves.



S_1 and S_2 are two very narrow slits. The single slit S ensures that the light leaving the slits S_1 and S_2 is coherent.

(c) **(i)** Define *coherent*. (1)

(ii) Explain why the slits S_1 and S_2 need to be very narrow. (2)

The point O on the diagram is equidistant from S_1 and S_2 and there is maximum constructive interference at point P on the screen. There are no other points of maximum interference between O and P.

(d) **(i)** State the condition necessary for there to be maximum constructive interference at the point P. (1)

(ii) On the axes below, draw a graph to show the variation of intensity of light on the screen between the points O and P.

(2)

(e) In this particular arrangement, the distance between the double slit and the screen is 1.50 m and the separation of S_1 and S_2 is 3.00×10^{-3} m.
The distance OP is 0.25 mm. Determine the wavelength of the light. (2)
Please note that part **(e)** is beyond the requirements of the syllabus.

(Total 14 marks)

5 This question is about wave properties and interference.

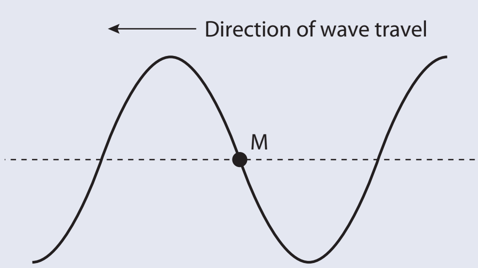
The diagram below represents the direction of oscillation of a disturbance that gives rise to a wave.

(a) By redrawing the diagram, add arrows to show the direction of wave energy transfer to illustrate the difference between

(i) a transverse wave and (1)

(ii) a longitudinal wave. (1)

A wave travels along a stretched string. The diagram below shows the variation with distance along the string of the displacement of the string at a particular instant in time. A small marker is attached to the string at the point labelled M. The undisturbed position of the string is shown as a dotted line.



(b) On the diagram above

(i) draw an arrow to indicate the direction in which the marker is moving. (1)

(ii) indicate, with the letter A, the amplitude of the wave. (1)

(iii) indicate, with the letter λ, the wavelength of the wave. (1)

(iv) draw the displacement of the string a time $\frac{T}{4}$ later, where T is the period of oscillation of the wave. Indicate, with the letter N, the new position of the marker. (2)

The wavelength of the wave is 5.0 cm and its speed is 10 cm s^{-1}.

(c) Determine

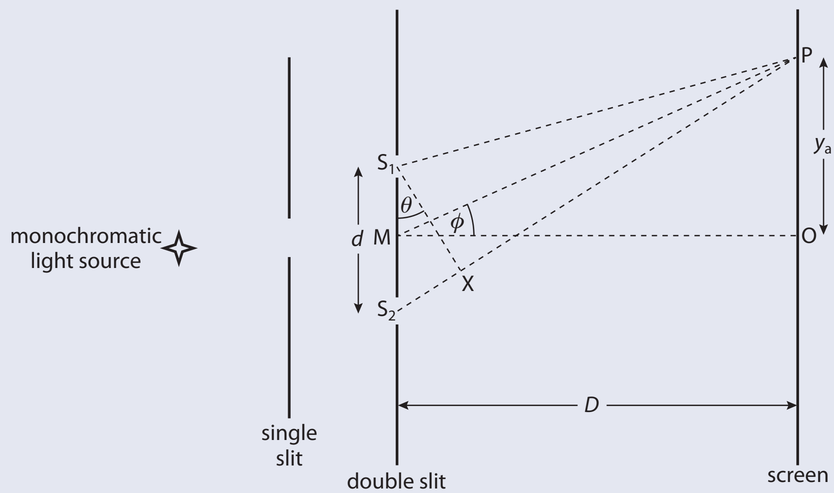
(i) the frequency of the wave. (1)

(ii) how far the wave has moved in $\frac{T}{4}$ s.

Interference of waves

(d) By reference to the principle of superposition, explain what is meant by constructive interference. (4)

The diagram below (not drawn to scale) shows an arrangement for observing the interference pattern produced by the light from two narrow slits S_1 and S_2.



The distance S_1S_2 is d, the distance between the double slit and screen is D and $D \gg d$ such that the angles θ and ϕ shown on the diagram are small. M is the mid-point of S_1S_2 and it is observed that there is a bright fringe at point P on the screen, a distance y_n from point O on the screen. Light from S_2 travels a distance S_2X further to point P than light from S_1

(e) **(i)** State the condition in terms of the distance S_2X and the wavelength of the light λ, for there to be a bright fringe at P. (2)

(ii) Deduce an expression for θ in terms of S_2X and d. (2)

(iii) Deduce an expression for ϕ in terms of D and y_n. (1)

For a particular arrangement, the separation of the slits is 1.40 mm and the distance from the slits to the screen is 1.50 m. The distance y_n is the distance of the eighth bright fringe from O and the angle $\theta = 2.70 \times 10^{-3}$ rad.

(f) Use your answers to **(e)** to determine

(i) the wavelength of the light (2)

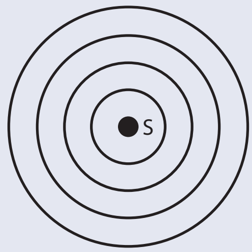
(ii) the separation of the fringes on the screen. (3)

Please note that part **(f)** is beyond the requirements of the syllabus.

(Total 24 marks)

6 This question is about the Doppler effect.

The diagram opposite shows wavefronts produced by a stationary wave source S. The spacing of the wavefronts is equal to the wavelength of the waves. The wavefronts travel with speed V.



(a) The source S now moves to the right with speed $\frac{1}{2}V$. Draw **four** successive wavefronts to show the pattern of waves produced by the moving source. (3)

(b) Derive the Doppler formula for the observed frequency f_0 of a sound source, as heard by a stationary observer, when the source approaches the stationary observer with speed v. The speed of sound is V and the frequency of the sound emitted by the source is f. (3)

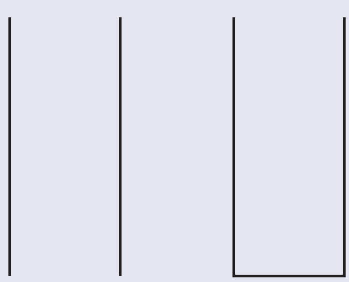
The Sun rotates about its centre. The light from one edge of the Sun, as seen by a stationary observer, shows a Doppler shift of 0.004 nm for light of wavelength 600.000 nm.

(c) Assuming that the Doppler formula for sound may be used for light, estimate the linear speed of a point on the surface of the Sun due to its rotation. (3)

(Total 9 marks)

7 This question is about standing waves in pipes.

The diagram below shows two pipes of the same length. Pipe A is open at both ends and pipe B is closed at one end.



(a) **(i)** On the diagrams above, draw lines to represent the waveforms of the fundamental (first harmonic) resonant note for each pipe. (2)

(ii) On each diagram, label the position of the nodes with the letter N and the position of the antinodes with the letter A. (2)

The frequency of the fundamental note for pipe A is 512 Hz.

(b) **(i)** Calculate the length of the pipe A. (Speed of sound in air $= 325\ \text{m s}^{-1}$) (3)

(ii) Suggest why organ pipes designed to emit low frequency fundamental notes (*e.g.* frequency $\approx$ 32 Hz) are often closed at one end. (2)

(Total 9 marks)

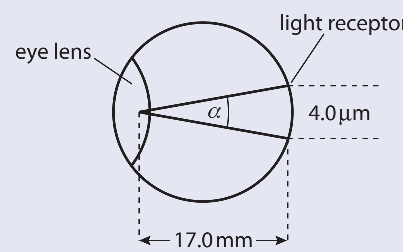
8 This question is about optical resolution.

(a) Light from a point source is brought to a focus by a convex lens. The lens does not cause spherical or chromatic aberration.

(i) State why the image of the point source will not be a point image. (1)

(ii) Describe the appearance of the image. (2)

Two light receptors at the back of the eye are 4.0 μm apart. The distance of the receptors from the convex lens at the front of the eye is 17.0 mm, as shown below.



Light of wavelength 550 nm from two point objects enters the eye. The centres of the images of the two objects are focused on the light receptors.

(b) **(i)** Calculate the angle α in radians subtended by the two receptors at the centre of the eye lens. (2)

(ii) Use the Rayleigh criterion to calculate the diameter of the pupil of the eye so that the two images are just resolved. (2)

(ii) With reference to your answer in **(i)**, suggest why the film appears to be coloured. (2)

(Total 9 marks)

SL Option A

Note that additional material for SL Option A (The eye and sight) is in the Appendix (page 577) as this is not required study for HL students.

5 Electrical currents

5.1 Electric potential difference, current and resistance

Assessment statements

5.1.1 Define electric potential difference.

5.1.2 Determine the change in potential energy when a charge moves between two points at different potentials.

5.1.3 Define the electronvolt.

5.1.4 Solve problems involving electric potential difference.

5.1.5 Define electric current.

5.1.6 Define resistance.

5.1.7 Apply the equation for resistance in the form $R = \frac{\rho l}{A}$ where ρ is the resistivity of the material of the resistor.

5.1.8 State Ohm's law.

5.1.9 Compare ohmic and non-ohmic behaviour.

5.1.10 Derive and apply expressions for electrical power dissipation in resistors.

5.1.11 Solve problems involving potential difference, current and resistance.

All electrical appliances convert electrical energy into other forms.

In this chapter we will develop the theory behind electric circuits, but first we need to understand how the individual components work.

The battery

All electrical devices convert electrical energy into other forms; an MP3 player produces sound, a torch produces light and an electric motor produces mechanical energy. In cordless devices (ones that don't plug into the mains) the source of this energy is the battery. Inside the battery, chemicals react with each other to convert chemical energy into electrical energy.

When the light bulb is connected across the ends of the battery, it gets hot and starts to glow, giving out energy in the form of light and heat. This energy comes from the battery. To make the bulb light up, we must make a complete circuit. If this circuit is broken, the bulb goes out. It seems like something is flowing through the wire; we call the flow *current* and the thing that is flowing we call *charge*.

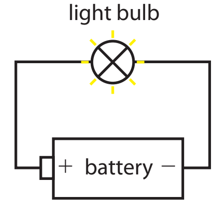


Figure 5.1 This simple circuit consists of a battery and a light bulb. The battery has two ends, one is positive (+) and the other is negative (−). To make a circuit, the bulb must be connected as shown .

W To build your own circuits and test the theories in this chapter, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on the Weblink 5.1.

Water analogy

To explain the observations, we can use an analogy.

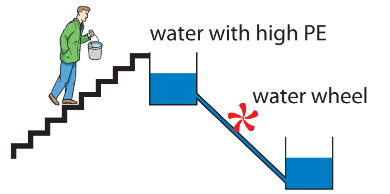


Figure 5.2 Water flow is similar to the flow of electricity.

Figure 5.2 shows a man carrying water up the stairs and pouring it into a tank. The water flows from this tank through a pipe into a lower one. As it flows, the water turns a water wheel. The water wheel turns a grindstone that rubs against a hard surface, producing heat.

Energy explanation

We can explain what is happening using the principle of conservation of energy. As the man carries the water up the stairs he does work on the water, increasing its potential energy(PE). When the water flows downhill into the lower tank, it loses PE. This energy is given to the water wheel, which turns the grindstone, producing heat. So we can say that the work done by the man is the same as the heat produced at the grindstone.

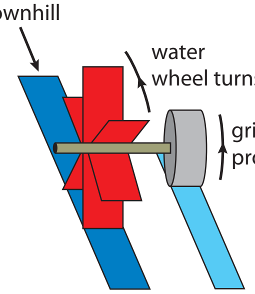


Figure 5.3 Energy changes.

Charge

Unlike with water, we cannot see charge flowing through the wire. We will find out later (Chapter 6) that there are two types of charge (+ and −) but to make things simple we will just consider + charge. The unit of charge is the coulomb (C).

Electrical potential energy

To make the water flow through the pipe, the man did work on the water to raise its potential energy. In the electrical circuit, the battery does work to increase the electrical potential energy of the charges. The charges on the + end of the battery have a higher potential energy than the ones on the − end, so charge flows from + to − outside the cell.

Current

Current is the flow of charge, and since charges at the + end of the battery have higher PE than those at the − end, the current flows from + to −.

The unit of current is the amp (A).

The amp is a scalar quantity.

$$\text{Current} = \frac{\text{charge}}{\text{time}}$$

1 amp = 1 coulomb per second.

The complete picture

The battery takes charges and puts them in a position of high potential at the + end. If a bulb is connected between the ends of the battery, this charge flows from the position of high potential to a position of low potential. The potential energy is converted to heat and light.

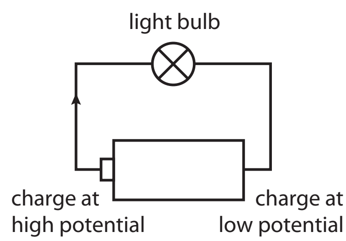


Figure 5.4 As the charge flows from + to −, its potential energy is converted into heat and light in the bulb.

Microscopic model

Metals allow charge to flow because they contain small negatively charged particles called *electrons*. Since the electrons are negative, they flow from low potential to high potential. Trying to imagine something flowing uphill is not very easy, so we will stick with positive charges flowing downhill. This is not a problem but we must always remember that electrons flow in the opposite direction to conventional current.

The electron is a small particle responsible for carrying charge in conductors.
Charge $= -1.6 \times 10^{-19}$ C
Electrons are negative so flow in the opposite direction to conventional current.

We consider current to flow in the opposite direction to the way that electrons actually flow. Does it matter that our model is actually completely backwards?

Energy transformation in a battery

Since energy is conserved, the energy given to the charges must come from somewhere – in fact, the battery converts chemical energy to electrical PE. The chemicals in the battery contain charges that are all mixed up. When these chemicals react, the charges are rearranged so that there are more + charges at the + end of the battery. After the reaction, the chemicals have less energy, as the chemical energy has been converted to electrical PE.

The PhET simulation 'Battery voltage' shows how the charges are rearranged. The little people represent chemical energy. To view this, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on the Weblink 5.2.

Potential difference

If we move a + charge from the negative end of the battery to the positive end, we need to do work. This is because the charge has a higher potential energy at the positive end than the negative end. The amount of work done per unit charge is defined as the potential difference (p.d.) between the plates or terminals.

The unit of p.d. is the volt. The volt is a scalar quantity.

1 volt = 1 joule per coulomb.

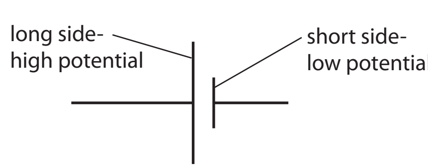


Figure 5.5 The symbol for a cell makes a lot of sense; the side that is at the highest potential has the longest line.

The p.d. between A and B is defined as the amount of work done per unit charge in taking a small +ve test charge from A to B.

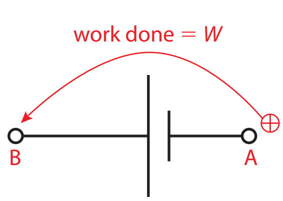


Figure 5.6 If an amount of work W is done in taking charge q from A to B then the p.d. (V) between A and B is given by the equation $V = W/q$.

Conductors and insulators

A *conductor* is a material that allows charge to flow through it. An *insulator* doesn't. All metals are good conductors of electricity, but plastics are insulators.

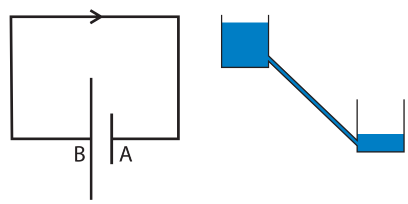
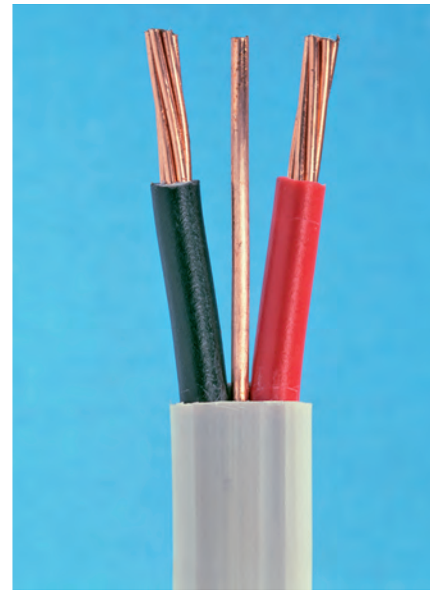


Figure 5.7 Current flows from B to A like water flowing downhill.



The metal core of the cable conducts electricity; the plastic cover is an insulator.

Worked examples

1. If a current of 10 A flows for 15 s, how much charge flows?
2. How much energy is lost when +5 C of charge flows from the +ve terminal to the −ve terminal of a 9 V battery?

Solutions

1 Current $I =$ charge per unit time $= \frac{Q}{t}$

$Q = It$

$= 10 \times 15$ C

$= 150$ C

2 Energy to move 1 C = 9 J

Energy to move 5 C = 5 × 9 J = 45 J

● **Examiner's hint:** From definition of p.d., the energy lost per unit charge moving from one terminal to the other is 9 J C^{-1}.

Resistance

The rate at which charge can flow through a conductor depends on the size and material of the conductor. A conductor that does not let much current flow for a given p.d. is said to have a high resistance.

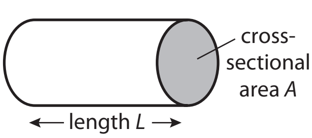


Figure 5.8 Dimensions of a conductor.

Resistance (R) is related to: cross-sectional area (A), length (L) and the material.

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

The constant of proportionality is called the *resistivity* (ρ).

So $$R = \rho\frac{L}{A}$$

The unit of resistance is the ohm (Ω).

This means that more current flows through a short fat conductor than a long thin one.

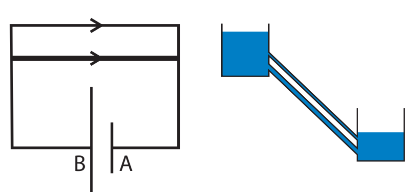


Figure 5.9 More current flows through the bottom branch than the top one.

Figure 5.10 The symbol for a resistor.

Resistivity

A resistor is a component with a known resistance. You can work out the resistance from the colours.

W To try the battery-resistance circuit, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 5.3.

By rearranging $R = \rho\frac{L}{A}$ we get $\rho = \frac{RA}{L}$ (units Ωm).

From this we can deduce that if the length of a sample of material is 1m and the cross-sectional area is $1\,\mathrm{m}^2$, then $\rho = R$. You can probably imagine that the resistance of such a large piece of metal will be very small – that's why the values of resistivity are so low (e.g. for copper $\rho = 1.72 \times 10^{-8}\ \Omega$m).

Microscopic model of resistance

As mentioned earlier, it is actually the negative charges called *free electrons* that move. As they move through the metal, they give energy to the metal atoms, causing the temperature of the metal to increase.

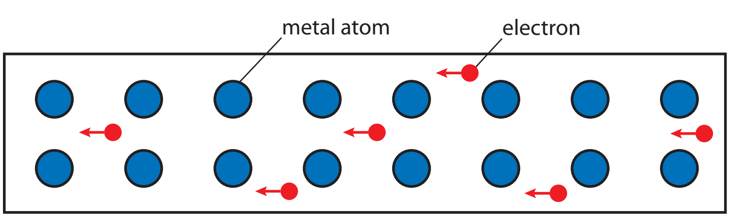


Figure 5.11 Electrons move through metal, giving energy to atoms.

Worked example

The resistivity of copper is $1.72 \times 10^{-8}\ \Omega$m. What is the resistance of a 1 m length of 2 mm diameter copper wire?

Solution

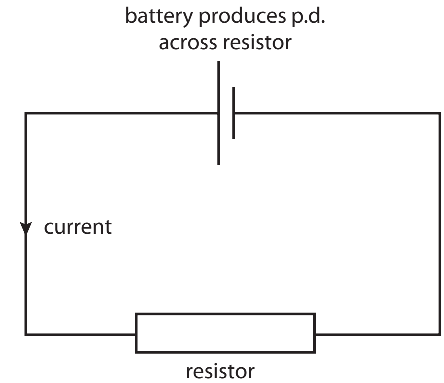
$$\text{Cross-sectional area} = \pi(0.001)^2$$
$$= 3.14 \times 10^{-6}\,\mathrm{m}^2$$
$$R = \rho\frac{L}{A}$$
$$= 1.72 \times 10^{-8} \times \frac{1}{3.14 \times 10^{-6}}\ \Omega$$
$$= 0.0055\ \Omega$$

● **Examiner's hint:**
Area $= \pi r^2$ where radius $= \frac{1}{2} \times 2\,\mathrm{mm} = 0.001\,\mathrm{m}$.

Ohm's law

Ohm's law relates the current flowing through a conductor with the potential difference across it.

Figure 5.12 We can now make a simple circuit with a battery and a resistor.



The current flowing through an ohmic conductor is directly proportional to the potential difference across it, provided temperature and other physical conditions remain constant.

Figure 5.13 Doubling p.d. doubles the current.

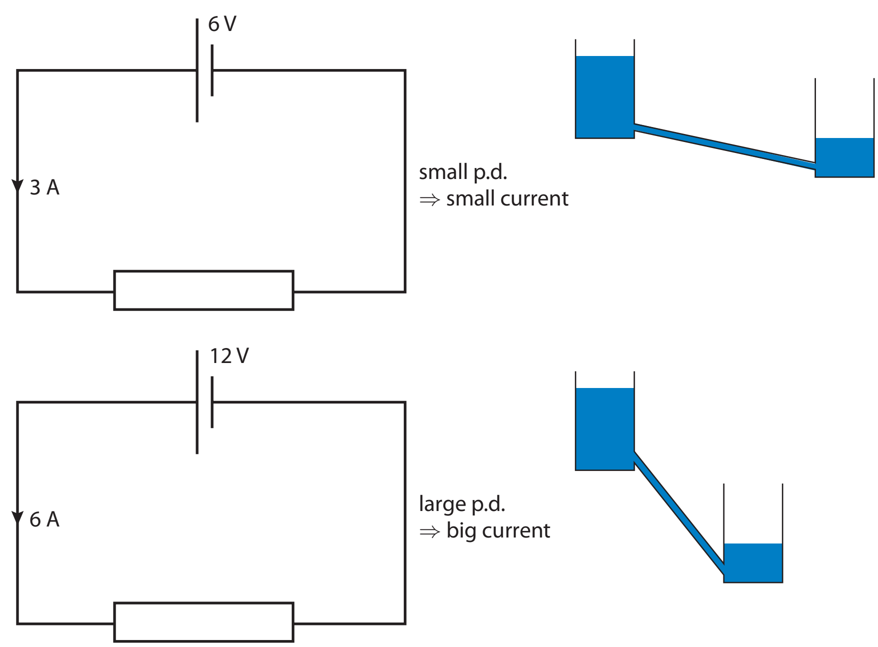
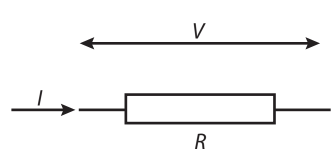


Figure 5.14 *V*, *I* and *R*



If the p.d. across a conductor is V and the current flowing through it is I, then according to Ohm's law:

$$V \propto I$$

The constant of proportionality is the resistance, R.

So: $$V = IR$$

Worked example

If the p.d. across a 3 Ω resistance is 9 V what current will flow?

Solution

$V = IR$ From Ohm's law.

$I = \frac{V}{R}$ Rearranging.

$I = \frac{9}{3}$ A

$= 3$ A

The water behind this dam has a high potential energy but because the hole is small only a small current of water can flow out. This is equivalent to a high resistance leading to a small current.

Graphical treatment

Ohmic conductor

Since $V \propto I$ for an ohmic conductor, a graph of I against V will be a straight line.

In this example, resistance $= \frac{V}{I} = 2\ \Omega$

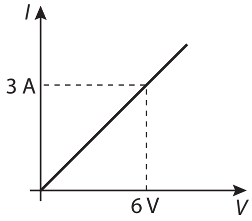


Figure 5.15 *I–V* for an ohmic conductor

Note: The resistance is found by taking the ratio $\frac{V}{I}$ – this is the same as $\frac{1}{\text{gradient}}$ for an ohmic conductor.

Non-ohmic conductors

Not all conductors obey Ohm's law. *I–V* graphs for these conductors will not be straight. A light bulb filament is an example of a non-ohmic conductor.

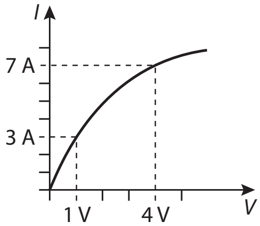


Figure 5.16 This *I–V* graph for a light bulb shows that the resistance has increased.

In this example, the resistance at the start is $\frac{1}{3}\ \Omega$ (0.33 Ω) and at the end it is $\frac{4}{7}\ \Omega$ (0.57 Ω).

The reason for this is that when the light bulb gets hot, the metal atoms vibrate more. This means that there are more collisions between the electrons and metal atoms, leading to an increase in resistance.

Why do we plot *V* on the *x*-axis? To produce an *I–V* graph, the current through the resistor is measured as the p.d. is changed. The variable that is changed is called the independent variable, the variable that changes is the dependent variable. The independent variable is generally plotted on the *x*-axis; that's why we plot ***I*** against *V*. So if an *I–V* graph gets steeper it means that the resistance is getting lower.

Exercises

1 If a p.d. of 9 V causes a current of 3 mA to flow through a wire, what is the resistance of the wire?

2 A current of 1 μA flows through a 300 kΩ resistor. What is the p.d. across the resistor?

3 If the p.d. across a 600 Ω resistor is 12 V, how much current flows?

4 Below is a table of the p.d. and current through a device called a thermistor.

p.d. / V	Current / A
1.0	0.01
10	0.1
25	1.0

Calculate the resistance at different potential differences.

5.2 Electric circuits

Assessment statements

5.2.1 Define electromotive force (emf).
5.2.2 Describe the concept of internal resistance.
5.2.3 Apply the equations for resistors in series and in parallel.
5.2.4 Draw circuit diagrams.
5.2.5 Describe the use of ideal ammeters and ideal voltmeters.
5.2.6 Describe a potential divider.
5.2.7 Explain the use of sensors in potential divider circuits.
5.2.8 Solve problems involving electric circuits.

The simple circuit – an energy view

We can now consider the energy changes that take place as current flows around the simple circuit.

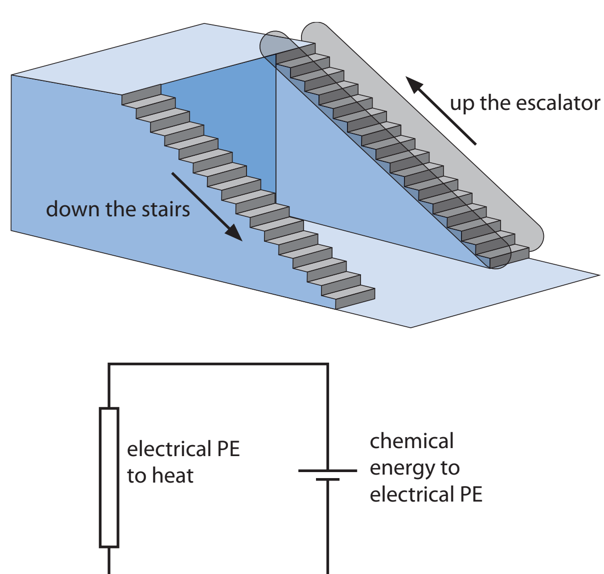


Figure 5.17 The shopping centre analogy. The energy changes of the charges that flow round an electric circuit are similar to the energy changes as you walk around a multi-storey shopping centre. The escalator gives you potential energy as it takes you up. When you come down the stairs your potential energy gets less. This energy is given to the stairs; as your feet hit each step, the step gets hot.

The cell uses chemical energy to place the charges in a position of high potential. As the charges flow through the resistor, this potential energy is converted to heat, so the resistor gets hot. These changes in energy are defined by the following terms:

- Emf (ε)

The emf of a cell is the amount of chemical energy converted to electrical energy, per unit charge. The unit is the volt (V).

- Potential difference (p.d. or V)

The p.d. across a resistance is the amount of electrical energy converted to heat, per unit charge. The unit is the volt (V).

Both of these quantities are energy per unit charge, but emf specifically applies to cells, batteries, generators and any other device that gives the charges potential energy. (One volt is the same as one joule per coulomb.)

To look at the simulations on the PhET website, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 5.4. 'Battery voltage' shows how the battery changes chemical energy to electrical potential energy by moving the charges. 'Battery-resistor current' shows how the electrical energy is converted to heat in the resistor. Note: these simulations show the movement of electrons, not conventional current.

Applying the law of conservation of energy

Since energy cannot be created or destroyed, the energy converted from chemical to electrical in the cell must be equal to the amount converted from electrical to heat in the resistor.

From Ohm's law, $V = IR$

Worked examples

1 If the emf of a battery is 9 V, how much energy is converted from chemical to electrical when 2 C of charge flow?

2 What is the p.d. across a resistor if 24 J of heat are produced when a current of 2 A flows through it for 10 s?

Solutions

1 emf = energy converted from chemical to electrical per unit charge.
So energy converted $= 2 \times 9$ J
$= 18$ J

2 2 A for 10 s $= 2 \times 10$ C $= 20$ C of charge.
If 20 C of charge flows, then the energy per unit charge $= \frac{24}{20}$ V $= 1.2$ V

Examiner's hint: If the amount of charge = 2 C then energy converted = 2 × emf

Examiner's hint: Current is charge flowing per unit time.

Examiner's hint: From the definition, p.d. = energy converted from electrical to heat per unit charge.

Internal resistance of cells

All cells are made of materials that have resistance. The resistance of the cell is called the *internal resistance*.

If a cell with internal resistance is connected to a resistor, current will flow from the cell. As current flows through the internal resistance, some energy is converted from electrical to heat inside the cell (so the cell gets hot). This means that there is less energy to be converted to heat in the resistor. The p.d. across the resistor is therefore less than the emf of the cell.

The shopping centre analogy is an example of how we use models to visualize things that we can't see. It is much easier to understand the changes in height as you walk around the shops than the changes of energy as electrons flow around a circuit.

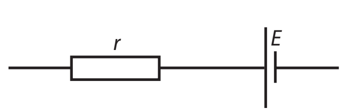
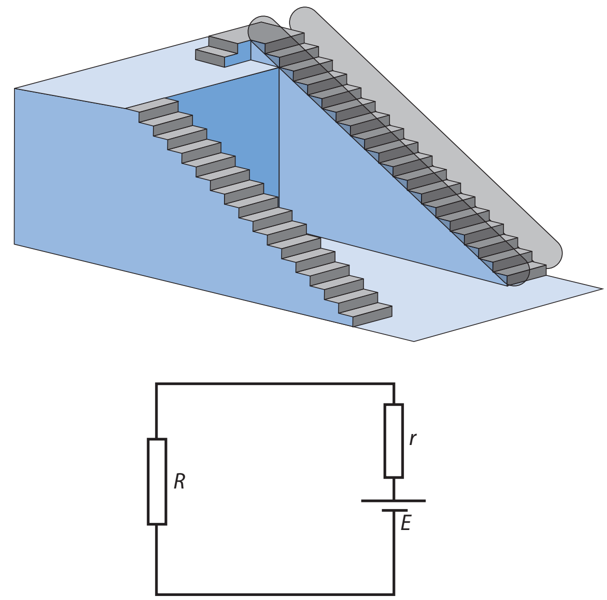


Figure 5.18 The symbol for a cell includes a resistor r next to the cell to show internal resistance.

Figure 5.19 Internal resistance acts like a small step down.



Shopping centre analogy
In the shopping centre analogy you can see that the elevator still lifts you up to the same level but you must come down a few steps as soon as you step off it.
In the electrical circuit the emf is still the same but the charges lose some energy before they leave the battery.

Applying Ohm's law to the internal resistance, the p.d. across it will be Ir.

From the law of conservation of energy, when a certain charge flows, the amount of energy converted from chemical to electrical equals the amount converted from electrical to heat.

$$\varepsilon = IR + Ir$$

Rearranging this formula, we can get an equation for the current from the battery.

$$I = \frac{\varepsilon}{R + r}$$

If the resistance connected to the battery is very small, then the current will be large. This means that most of the electrical energy is converted to heat inside the battery so the battery gets very hot. This is why you shouldn't connect a wire between the ends of a battery. You can try this with the PhET circuit construction kit, but don't try it with a real battery.

Worked example

A battery of emf 9 V with an internal resistance 1 Ω is connected to a 2 Ω resistor, as shown in Figure 5.20.

How much current will flow?

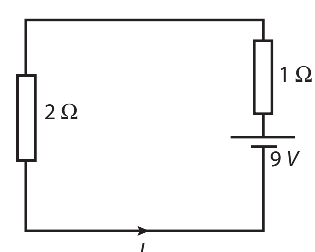


Figure 5.20 Always start by drawing a circuit showing the quantities you know and labelling the ones you want to find.

Solution

$$I = \frac{\varepsilon}{R + r}$$

$$I = \frac{9}{2 + 1}\,\text{A}$$

$$= 3\,\text{A}$$

What is the p.d. across the 2 Ω resistor?

$$V = IR$$

$$V = 3 \times 2\,\text{V}$$

$$= 6\,\text{V}$$

● **Examiner's hint:** The current through the resistor is 3 A.

Exercises

5 A current of 0.5 A flows when a battery of emf 6 V is connected to an 11 Ω resistor. What is the internal resistance of the battery?

6 A 12 V battery with internal resistance 1 Ω is connected to a 23 Ω resistor. What is the p.d. across the 23 Ω resistor?

Electrical power

Electrical power is the rate at which energy is changed from one form to another.

Power delivered

In a perfect battery, the power is the amount of chemical energy converted to electrical energy per unit time.

If the emf of a battery is ε, then if a charge Q flows, the amount of energy converted from chemical to electrical is εQ.

If this charge flows in a time t then the power delivered $= \dfrac{\varepsilon Q}{t}$

But $\dfrac{Q}{t}$ = the current, I

So the power delivered $= \varepsilon I$

In a real battery, the actual power delivered will be a bit less, since there will be some power dissipated in the internal resistance.

Kilowatt-hours

The electrical energy used at home is measured in kilowatt-hours. This is the amount of energy used by a 1 kilowatt heater switched on for 1 hour. When the electricity bill comes, you have to pay for each kilowatt-hour of energy that you have used.

Since $1\,\text{W} = 1\,\text{J s}^{-1}$

1 kilowatt-hour $= 1000 \times 60 \times 60\,\text{J}$

$1\,\text{kWh} = 3.6 \times 10^6\,\text{J}$

Power dissipated

The power dissipated in the resistor is the amount of electrical energy converted to heat per unit time.

Consider a resistance R with a p.d. V across it. If a charge q flows in time t then the current, $I = \dfrac{q}{t}$

The p.d., V, is defined as the energy converted to heat per unit charge, so the energy converted to heat in this case $= Vq$

Power is the energy used per unit time, so $P = \dfrac{Vq}{t}$

but $\dfrac{q}{t} = I$, so $P = VI$

Worked examples

1 If a current of 2 A flows through a resistor that has a p.d. of 4 V across it, how much power is dissipated?

2 What power will be dissipated when a current of 4 A flows through a resistance of 55 Ω?

Solutions

1 $P = VI$ where $V = 4\,\text{V}$ and $I = 2\,\text{A}$

$P = 4 \times 2\,\text{W}$

$= 8\,\text{W}$

2 Using Ohm's law $V = IR$ — First we need to find out the p.d. across the resistor

$= 4 \times 55\,\text{V}$

$= 220\,\text{V}$

$P = VI$

$= 220 \times 4\,\text{W}$

$= 880\,\text{W}$

Alternative ways of writing $P = VI$

In Example 2, we had to calculate the p.d. before finding the power. It would be convenient if we could solve this problem in one step. We can write alternative forms of the equation by substituting for I and V from Ohm's law.

We have shown that power $P = VI$

But from Ohm's law $V = IR$

If we substitute for V we get $P = IR \times I = I^2R$

We can also substitute for $I = \dfrac{V}{R}$

Power $= V \times I = \dfrac{V \times V}{R} = \dfrac{V^2}{R}$

Power
$P = VI$
$P = I^2R$
$P = V^2/R$

Exercises

7 5 A flows through a 20 Ω resistor.
(a) How much electrical energy is converted to heat per second?
(b) If the current flows for one minute, how much energy is released?

8 If a battery has an internal resistance of 0.5 Ω, how much power will be dissipated in the battery when 0.25 A flows?

9 A current of 0.5 A flows from a battery of emf 9 V. If the power delivered is 4 W, how much power is dissipated in the internal resistance?

Electric kettle (water boiler)

An electric kettle transfers the heat produced when current flows through a wire element to the water inside the kettle.

Worked example

A current of 3 A flows through an electric kettle connected to the 220 V mains. What is the power of the kettle and how long will it take to boil 1 litre of water?

Solution

The power of the kettle $= VI = 220 \times 3 = 660$ W

To calculate energy needed to boil the water, we use the formula

heat required = mass × specific heat capacity × temperature change.

The specific heat capacity of water is 4180 J kg^{-1} °C^{-1}

The mass of 1 litre of water is 1 kg, so if we assume that the water was at room temperature, 20°C, then to raise it to 100°C the energy required is:
$1 \times 4180 \times 80 = 334\,400$ J

$$\text{power} = \text{energy/time, so the time taken} = \frac{\text{energy}}{\text{power}}$$
$$= \frac{334\,400}{660}$$
$$\text{Time} = 506.67 \text{ s}$$
$$= 8 \text{ minutes } 27 \text{ seconds}$$

The light bulb

If the power dissipated in a wire is large, then a lot of heat is produced per second. When heat is added quickly, the wire doesn't have time to lose this heat to the surroundings. The result is that the temperature of the wire increases, and if the temperature is high enough, the wire will begin to glow, giving out light. Only about 10% of the energy dissipated in a light bulb is converted to light - the rest is heat.

Fluorescent tubes are much more efficient than incandescent light bulbs, converting most of the electrical energy to light.

The electric motor

A motor converts electrical energy to mechanical energy; this could be in the form of potential energy, if something is lifted by the motor, or kinetic energy, if the motor is accelerating something like a car.

Worked example

An electric motor is used to lift 10 kg through 3 m in 5 seconds. If the p.d. across the motor is 12 V, how much current flows (assuming no energy is lost)?

Solution

$$\text{Work done by the motor} = mgh = 10 \times 10 \times 3\ \text{J} = 300\ \text{J}$$

$$\text{Power} = \frac{\text{work done}}{\text{time}} = \frac{300}{5}\ \text{W} = 60\ \text{W}$$

Electrical power, $P = IV$ so $I = P/V$

$$= \frac{60}{12}\ \text{A} = \mathbf{5\ A}$$

Exercises

10 An electric car of mass 1000 kg uses twenty-five 12 V batteries connected together to create a p.d. of 300 V. The car accelerates from rest to a speed of 30 m s^{-1} in 12 seconds.
- **(a)** What is the final kinetic energy of the car?
- **(b)** What is the power of the car?
- **(c)** How much electrical current flows from the battery?

What assumptions have you made in calculating **a)** to **c)**?

11 A light bulb for use with the 220 V mains is rated at 100 W.
- **(a)** What current will flow through the bulb?
- **(b)** If the bulb converts 20% of the energy to light, how much light energy is produced per second?

12 A 1 kW electric heater is connected to the 220 V mains and left on for 5 hours.
- **(a)** How much current will flow through the heater?
- **(b)** How much energy will the heater release?

Combinations of components

In practical situations, resistors and cells are often joined together in combinations e.g. fairy lights, flashlight batteries.

There are many ways of connecting a number of components – we will consider two simple arrangements, series and parallel.

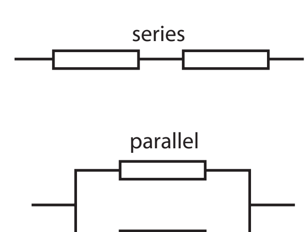
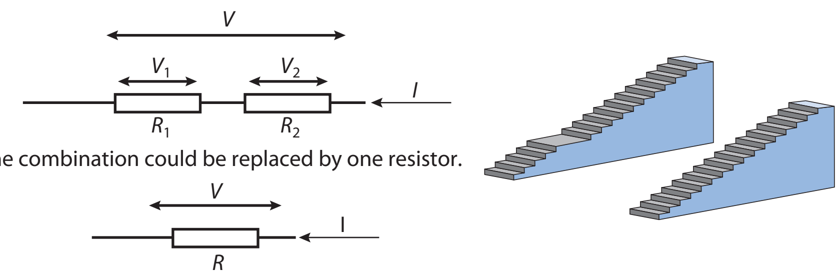


Figure 5.21 Two simple combinations of resistors.

Resistors in series

In a series circuit the same current flows through each resistor.

Figure 5.22 Two resistors in series are similar to two flights of stairs.



These coloured lights are connected in series – if you take one out they all go out.

Applying the law of conservation of energy, the p.d. across R_1 plus the p.d. across R_2 must be equal to the p.d. across the combination.

$$V_1 + V_2 = V$$

Applying Ohm's law to each resistor: $IR_1 + IR_2 = IR$

Dividing by I: $R_1 + R_2 = R$

Worked example

What is the total resistance of a 4Ω and an 8Ω resistor in series?

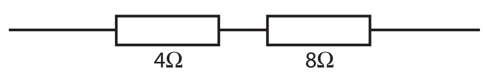


Figure 5.23

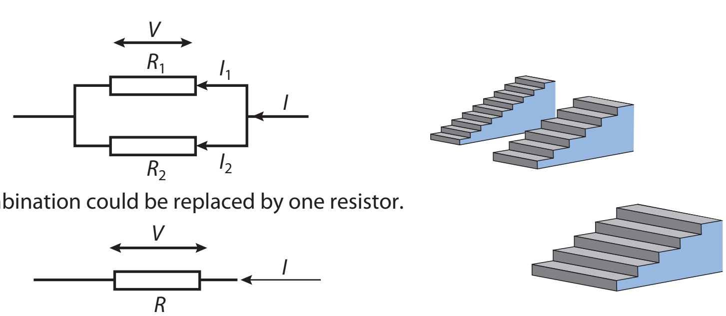
Solution

Total resistance $= R_1 + R_2$
$= 4 + 8\ \Omega$
$= 12\ \Omega$

Resistors in parallel

In a parallel circuit the current splits in two.

Figure 5.24 Resistors in parallel are similar to stairs side by side.



Applying the law of conservation of charge, we know that the current going into a junction must equal the current coming out.

$$I = I_1 + I_2$$

Applying Ohm's law to each resistor gives: $\frac{V}{R} = \frac{V}{R_1} + \frac{V}{R_2}$

Dividing by V: $\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}$

Worked examples

1 What is the total resistance of a 4 Ω and an 8 Ω resistor in parallel?

2 What is the total resistance of two 8 Ω resistors in parallel?

Solutions

1 Using $\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}$

$$\frac{1}{R} = \frac{1}{4} + \frac{1}{8}$$

$$= \frac{2+1}{8} = \frac{3}{8}$$

so $R = \frac{8}{3}\,\Omega$

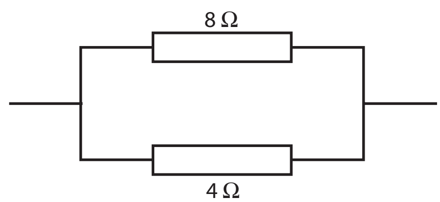


Figure 5.25

2 Using $\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}$

$$\frac{1}{R} = \frac{1}{8} + \frac{1}{8} = \frac{2}{8}$$

$$R = \frac{8}{2}\,\Omega$$

$$= 4\,\Omega$$

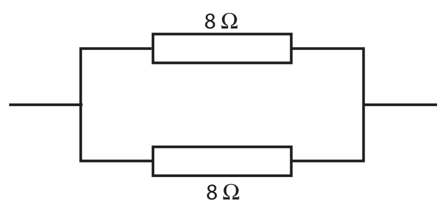


Figure 5.26 The total resistance of two equal resistors in parallel is half the resistance of one of them.

Multiple combinations

When there are many resistors, the total is found by splitting the circuit into small units of parallel and series resistors; for example in the circuit shown in Figure 5.27.

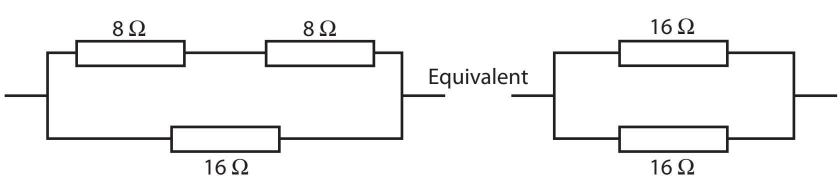


Figure 5.27 Equivalent circuits.

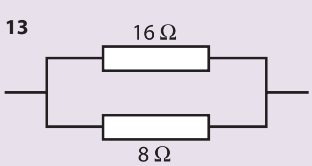
The two 8 Ω resistors at the top are equivalent to one 16 Ω resistor.

This gives two16 Ω resistors in parallel, a total of 8 Ω.

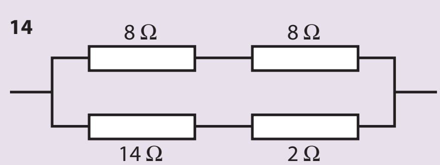
Exercises

Calculate the total resistance for the circuits in Figure 5.28.

13



14



15

16

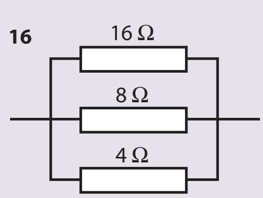
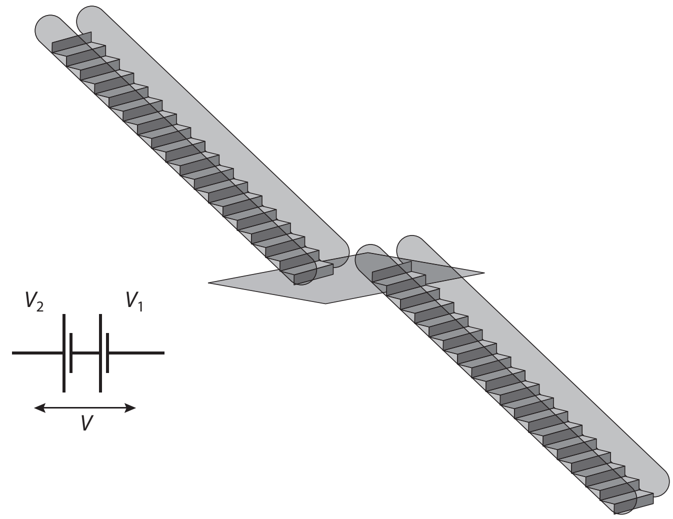


Figure 5.28

Cells in series

Cells are very often added in series to produce a larger p.d.

Figure 5.29 Cells in series are similar to two flights of escalators.



The p.d. of cells in series simply add

$$V = V_1 + V_2$$

Worked example

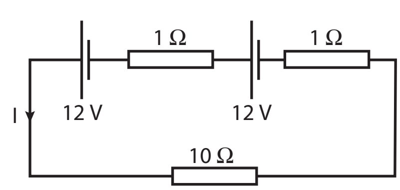
The three resistors are in series (it looks like the 10 Ω is in parallel with the 1 Ω resistors, but they are in series because the same current flows through them all).

Two 12 V batteries are connected in series to a 10 Ω resistor. If each battery has an internal resistance of 1 Ω, how much current will flow?

Solution

The total p.d. for two batteries in series = 12 + 12 = 24 V

Figure 5.30 The circuit diagram.



Total resistance = 1 + 1 + 10 = 12 Ω.

Applying Ohm's law,

$$\begin{aligned} I &= \frac{V}{R} \\ &= \frac{24}{12}\text{ A} \\ &= 2\text{ A} \end{aligned}$$

Electrical measurement

Measurement of potential difference

The p.d. can be measured using a voltmeter. There are two main types of voltmeter, digital and analogue.

The p.d. is the difference in potential between two points. To measure the p.d. between A and B, one lead of the voltmeter must be connected to A, the other to B.

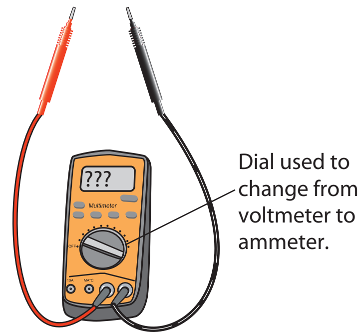


Figure 5.31 A multimeter is a common instrument that can measure both p.d. and current. It can also measure resistance.

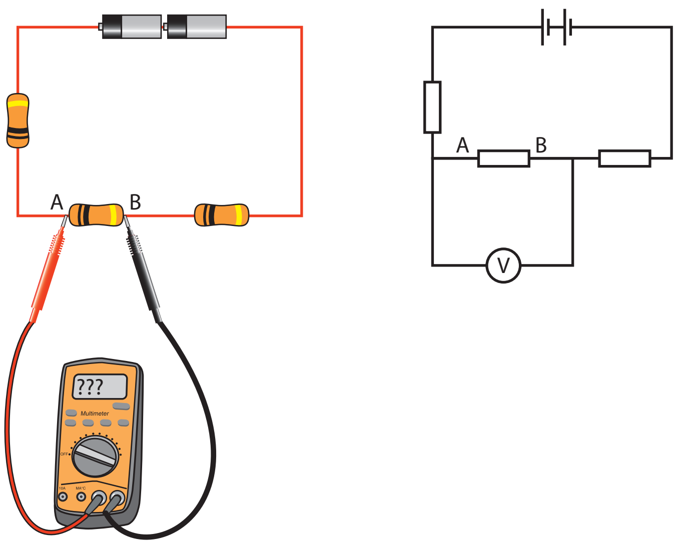


Figure 5.32 A voltmeter is connected from A to B.

An ideal voltmeter has infinitely high resistance so that it does not take any current from the circuit.

Measurement of current

To measure the current flowing through a resistor, the ammeter must be connected so that the same current will flow through the ammeter as flows through the resistor. This means disconnecting one of the wires and connecting the ammeter.

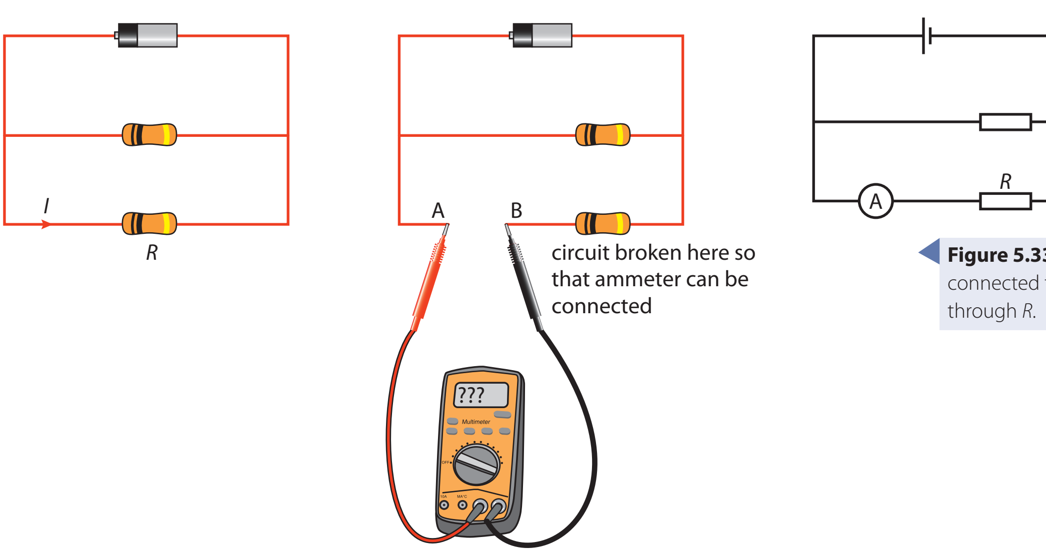


Figure 5.33 The ammeter is connected to measure the current through *R*.

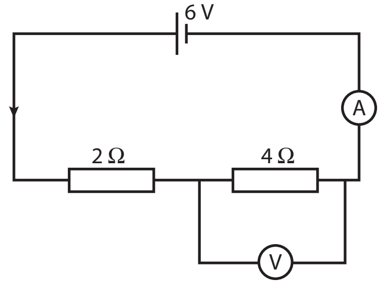
An ideal ammeter has zero resistance so that it doesn't change the current in the circuit.

Worked example

Calculate the current and potential difference measured by the meters in the circuit in Figure 5.34. Assume the battery has no internal resistance and that the meters are ideal.

Solution

Figure 5.34



Total resistance $= 2 + 4\,\Omega$

$= 6\,\Omega$ — The two resistors are in series.

$I = \dfrac{V}{R}$ — Applying Ohm's law, to find the current.

$= \dfrac{6}{6}\,\text{A}$

$= 1\text{A}$

So ammeter reading $= 1\text{A}$

$V = IR$ — Applying Ohm's law to the $4\,\Omega$ resistor.

$= 1 \times 4\,\text{V}$

$= 4\,\text{V}$

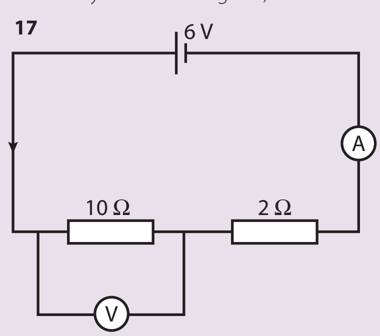
So the voltmeter reads 4 V.

Exercises

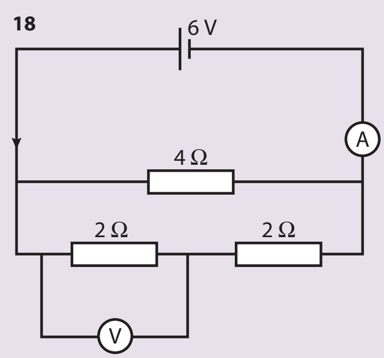
Find the ammeter and voltmeter readings in the circuits in Figure 5.35. All meters are ideal and the batteries have no internal resistance. (You can build them in the PhET 'circuit construction kit' to see whether your answers agree.)

Figure 5.35

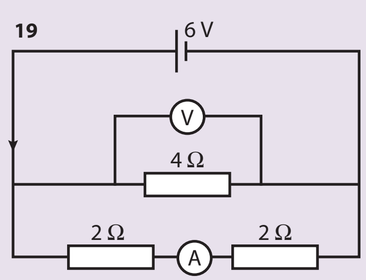
17



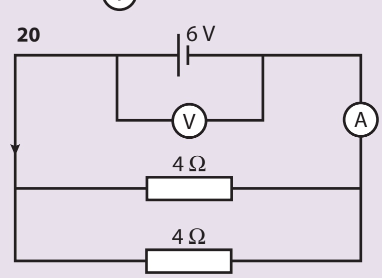
18



19



20



Electrical sensors

An electrical sensor is a device whose electrical properties change with changing physical conditions.

Thermistor

A thermistor is made from a semiconducting material whose resistance decreases as temperature increases. As the thermistor gets hotter, more charge carriers are released, so the current can flow more easily.

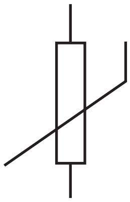


Figure 5.36 A thermistor.

Light sensor (LDR)

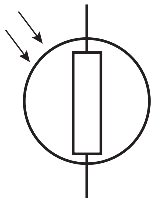


Figure 5.37 An LDR.

A light sensor or light-dependent resistor (LDR) is also a semiconducting device, but, unlike the thermistor, it is light that releases more charge carriers, resulting in a lower resistance.

Strain gauge

A strain gauge is a thin metal wire. If it is stretched, its length increases and its cross-sectional area gets smaller. This results in an increase in resistance.

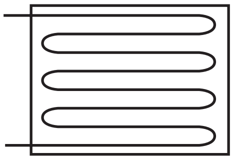


Figure 5.38 A strain gauge.

Use of sensors

The resistance of all three sensors varies with some physical property. However, it would be much more useful if the devices gave a changing p.d. rather than a changing resistance. To convert the changing resistance to a changing p.d., we use a potential divider.

The potential divider

The battery creates a p.d. across the resistors equal to V_{in}

From Ohm's law we know that the current in the circuit $I = \frac{V}{R}$

Since the total resistance is $R_1 + R_2$

The current $I = \frac{V_{in}}{R_1 + R_2}$ Equation (1)

The p.d. across $R_2 = V_{out}$

Applying Ohm's law to R_2 gives

$V_{out} = IR_2$

Substituting from equation (1) gives

$$V_{out} = V_{in}\frac{R_2}{R_1 + R_2}$$

This is the potential divider formula.

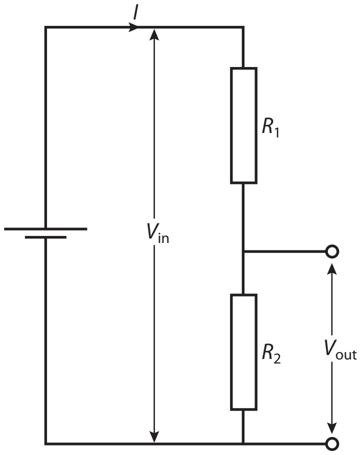
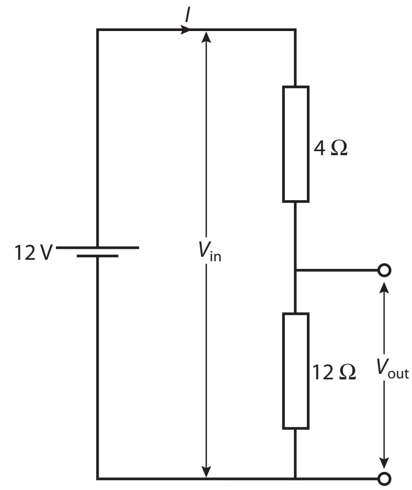


Figure 5.39 A potential divider circuit consists of two series resistors.

Worked example

Calculate the output voltage for the potential divider in Figure 5.40.

Figure 5.40



Solution

Using the potential divider formula

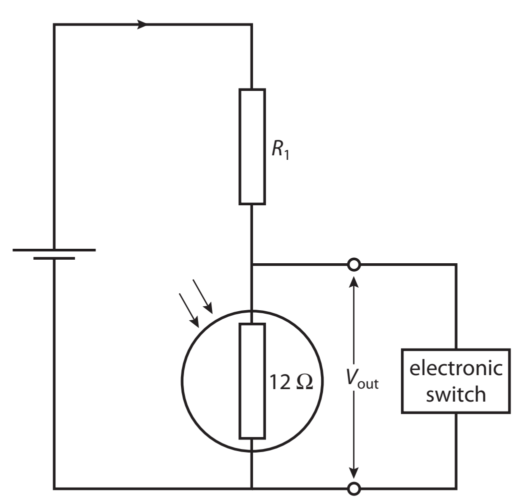
$$V_{out} = V_{in}\frac{R_2}{R_1 + R_2}$$

$$V_{out} = 12\frac{12}{4 + 12} = 9\,\text{V}$$

Using the potential divider with sensors

Automatic light switch

Figure 5.41 An LDR and a potential divider can be used to operate an automatic light switch.



When light stops shining on the LDR, its resistance increases, resulting in an increase in V_{out}. The increase in V_{out} in turn activates the electronic switch that puts on the lights. The electronic switch needs a minimum p.d. to activate it, so it doesn't switch on the lights until V_{out} is big enough.

The important point here is that the electronic switch needs a p.d. to activate it, hence the need for a potential divider.

Worked example

The battery in Figure 5.41 has an emf of 12 V and no internal resistance. The p.d. required to activate the switch is 5 V. Find the value of R_1 that will cause the lights to turn on when the resistance of the LDR rises to 200 kΩ.

Solution

$V_{in} = 12\,\text{V}$

$R_2 = 200\,\text{k}\Omega$

$V_{out} = 5\,\text{V}$

Rearranging the potential divider equation $V_{out} = V_{in}\dfrac{R_2}{R_1 + R_2}$

gives $R_1 = R_2\dfrac{V_{in} - V_{out}}{V_{out}}$

$$= 200\,\text{k}\Omega \times \frac{(12 - 5)}{5}$$

$$= 280\,\text{k}\Omega$$

The fire alarm

When the thermistor gets hot, its resistance decreases, resulting in an increased current through R_2, which in turn leads to an increase in V_{out}. The increase in V_{out} activates an electronic switch that rings a bell.

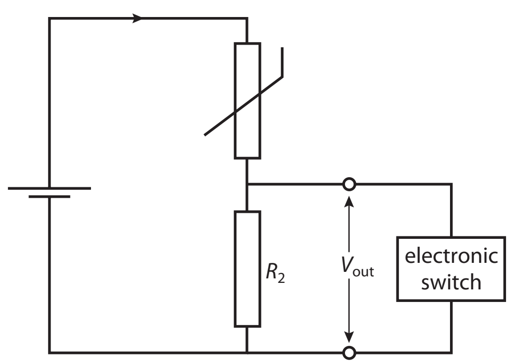


Figure 5.42 The fire alarm bell must ring if the thermistor gets hot.

Exercises

21 Assume that the circuit in Figure 5.42 has a 12 V battery and a switch that activates when the p.d. is 5 V. Calculate the value of R_2, if the bell rings when the resistance of the thermistor drops to 1 kΩ?

Using a strain gauge

A strain gauge can be used to detect whether parts of a building are stretching. For example, a strain gauge stuck to the underside of a bridge will be stretched if the bridge bends when a heavy truck crosses it. If the strain gauge is connected in a potential divider circuit, the V_{out} can be used to measure how much the bridge stretches.

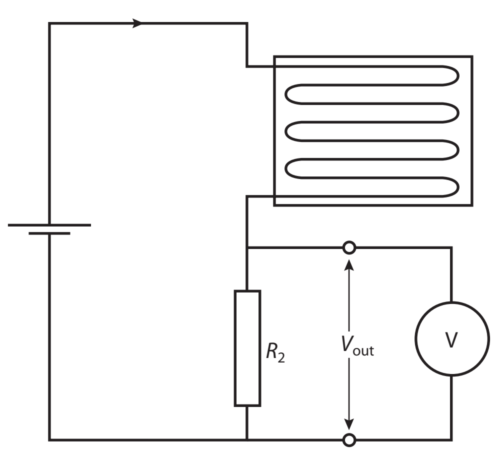


Figure 5.43 Strain gauge circuit.

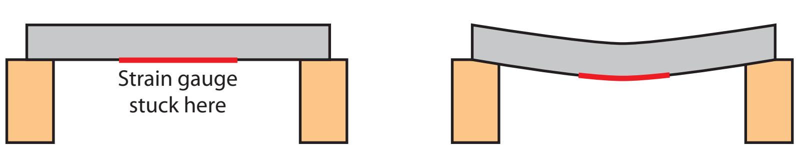
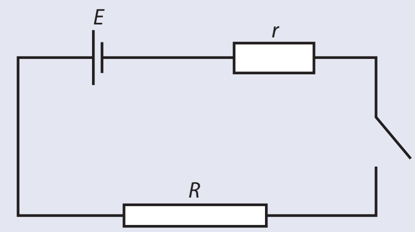


Figure 5.44 As the bridge bends, the strain gauge gets longer.

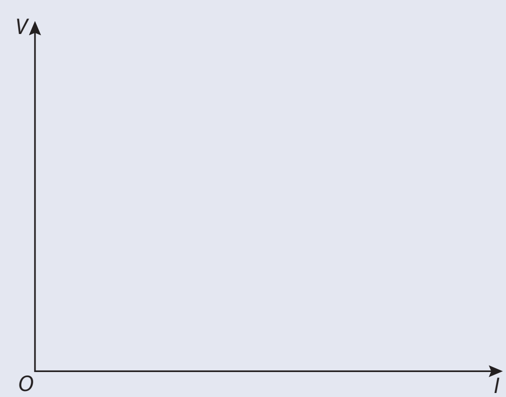
Practice questions

1 This question is about electrical energy and associated phenomena.
A cell of emf E and internal resistance r is connected in series with a resistor R, as shown below. The cell supplies 8.1×10^3 J of energy when 5.8×10^3 C of charge moves completely round the circuit. The current in the circuit is constant.



(i) Calculate the emf E of the cell. (2)
(ii) The resistor R has resistance 6.0 Ω. The potential difference between its terminals is 1.2 V. Determine the internal resistance r of the cell. (3)
(iii) Calculate the total energy transfer in the resistor R. (2)
(iv) Describe, in terms of a simple model of electrical conduction, the mechanism by which the energy transfer in the resistor R takes place. (5)

2 This question is about a filament lamp.
(a) On the axes below, draw a sketch graph to show the variation with potential difference V of the current I in a typical filament lamp (the I–V characteristic). (Note: this is a sketch graph; you do not need to add any values to the axes). (1)



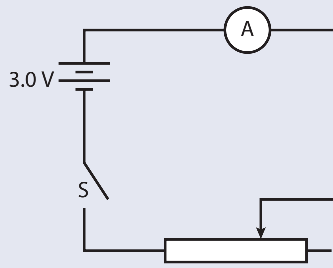
(b) (i) Explain how the resistance of the filament is determined from the graph. (1)
(ii) Explain whether the graph you have sketched indicates ohmic behaviour or non-ohmic behaviour. (1)

A filament lamp operates at maximum brightness when connected to a 6.0 V supply. At maximum brightness, the current in the filament is 120 mA.

(c) (i) Calculate the resistance of the filament when it is operating at maximum brightness. (1)
(ii) You have available a 24 V supply and a collection of resistors of a suitable power rating and with different values of resistance. Calculate the resistance of the resistor that is required to be connected in series with the supply such that the voltage across the filament lamp will be 6.0 V. (2)

3 This question is about electric circuits.

Susan sets up the circuit below in order to measure the current–voltage (I–V) characteristic of a small filament lamp.



The supply is a battery that has an emf of 3.0 V and the ammeter and voltmeter are considered to be ideal. The lamp is labelled by the manufacturer as '3 volts, 0.6 watts'.

(a) **(i)** Explain what information this labelling provides about the normal operation of the lamp. (2)

(ii) Calculate the current in the filament of the lamp when it is operating at normal brightness. (2)

Susan sets the variable resistor to its maximum value of resistance. She then closes the switch S and records the following readings.

Ammeter reading = 0.18 A Voltmeter reading = 0.60 V

She then sets the variable resistor to its zero value of resistance and records the following readings.

Ammeter reading = 0.20 A Voltmeter reading = 2.6 V

(b) **(i)** Explain why, by changing the value of the resistance of the variable resistance, the potential difference across the lamp cannot be reduced to zero or be increased to 3.0 V. (2)

(ii) Determine the internal resistance of the battery. (3)

(c) Calculate the resistance of the filament when the reading on the voltmeter is

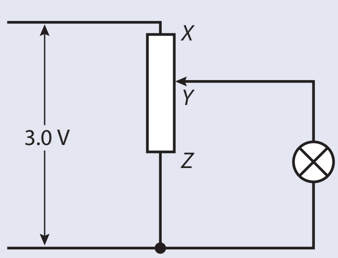
(i) 0.60 V. (1)

(ii) 2.6 V. (1)

(d) Explain why there is a difference between your answers to **(c) (i)** and **(c) (ii)**. (2)

(e) Using the axes as in question 2, draw a sketch-graph of the I–V characteristic of the filament of the lamp. (Note: this is a sketch-graph; you do not need to add any values to the axis.) (1)

The diagram below shows an alternative circuit for varying the potential difference across the lamp.

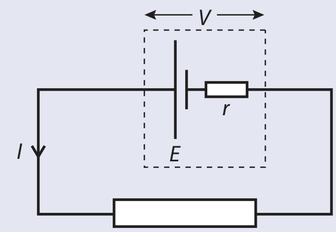


The potential divider XZ has a potential of 3.0 V across it. When the contact is at the position Y, the resistance of XY equals the resistance of YZ which equals 12 Ω. The resistance of the lamp is 4 Ω.

(f) Calculate the potential difference across the lamp. (4)

4 This question is about emf and internal resistance.

A dry cell has an emf E and internal resistance r and is connected to an external circuit. There is a current I in the circuit when the potential difference across the terminals of the cell is V.



(a) State expressions, in terms of E, V, r and I where appropriate, for

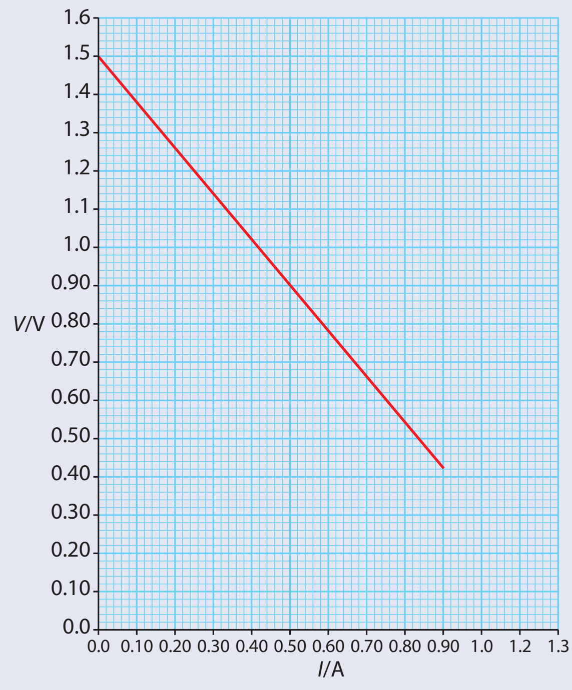
(i) the total power supplied by the cell. (1)

(ii) the power dissipated in the cell. (1)

(iii) the power dissipated in the external circuit. (1)

(b) Use your answers to **(a)** to derive a relationship between V, E, I and r. (2)

The graph below shows the variation of V with I for the dry cell.



(c) Draw the circuit that could be used to obtain the data from which the graph was plotted. (3)

(d) Use the graph, explaining your answers, to

(i) determine the emf E of the cell. (2)

(ii) determine the current in the external circuit when the resistance R of the external circuit is very small. (2)

(iii) deduce that the internal restance r of the cell is about 1.2 Ω. (1)

(e) The maximum power dissipated in the external circuit occurs when the resistance of the external circuit has the same value as the internal resistance of the cell. Calculate the maximum power dissipation in the external circuit. (3)

6 Fields and forces

6.1 Gravitational force and field

Assessment statements

6.1.1 State Newton's universal law of gravitation.
6.1.2 Define *gravitational field strength*.
6.1.3 Determine the gravitational field due to one or more point masses.
6.1.4 Derive an expression for gravitational field strength at the surface of a planet, assuming that all its mass is concentrated at its centre.
6.1.5 Solve problems involving gravitational forces and fields.

Gravitational force and field

We have all seen how an object falls to the ground when released. Newton was certainly not the first person to realize that an apple falls to the ground when dropped from a tree. However, he did recognize that the force that pulls the apple to the ground is the same as the force that holds the Earth in its orbit around the Sun; this was not obvious - after all, the apple moves in a straight line and the Earth moves in a circle. In this chapter we will see how these forces are connected.

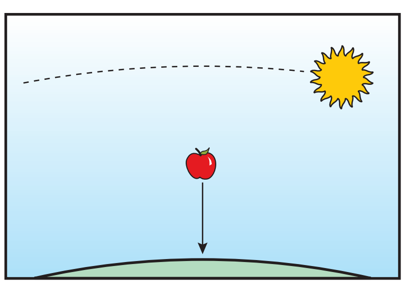


Figure 6.1 The apple drops and the Sun seems to move in a circle, but it is gravity that makes both things happen.

Newton's universal law of gravitation

Newton extended his ideas further to say that every single particle of mass in the universe exerts a force on every other particle of mass. In other words, everything in the universe is attracted to everything else. So there is a force between the end of your nose and a lump of rock on the Moon.

The modern equivalent of the apparatus used by Cavendish to measure *G* in 1798.

Was it reasonable for Newton to think that his law applied to the whole universe?

Newton's universal law of gravitation states that:

> every single point mass attracts every other point mass with a force that is directly proportional to the product of their masses and inversely proportional to the square of their separation.

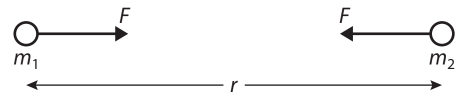


Figure 6.2 The gravitational force between two point masses.

If two point masses with mass m_1 and m_2 are separated by a distance r then the force, F, experienced by each will be given by:

$$F \propto \frac{m_1 m_2}{r^2}$$

The constant of proportionality is the universal gravitational constant G.

$$G = 6.6742 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$$

Therefore the equation is $\quad F = G\dfrac{m_1 m_2}{r^2}$

Spheres of mass

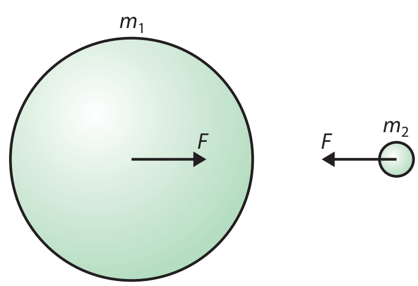


Figure 6.3 Forces between two spheres. Even though these bodies don't have the same mass, the force on them is the same size. This is due to Newton's third law – if mass m_1 exerts a force on mass m_2 then m_2 will exert an equal and opposite force on m_1.

By working out the total force between every particle of one sphere and every particle of another, Newton deduced that spheres of mass follow the same law, where the separation is the separation between their centres. Every object has a centre of mass where the gravity can be taken to act. In regularly-shaped bodies, this is the centre of the object.

How fast does the apple drop?

If we apply Newton's universal law to the apple on the surface of the Earth, we find that it will experience a force given by

$$F = G\frac{m_1 m_2}{r^2}$$

where:

m_1 = mass of the Earth = 5.97×10^{24} kg

m_2 = mass of the apple = 250 g

r = radius of the Earth = 6378 km (at the equator)

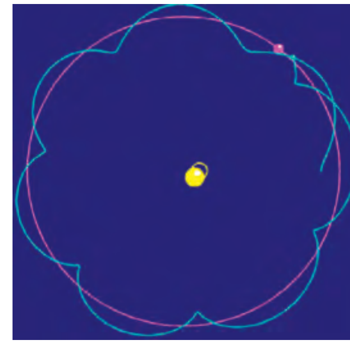
So $\quad F = 2.43$ N

From Newton's 2nd law we know that $F = ma$.

So the acceleration (a) of the apple $= \dfrac{2.43}{0.25}\,\mathrm{m\,s^{-2}}$

$a = 9.79\,\mathrm{m\,s^{-2}}$

This is very close to the average value for the acceleration of free fall on the Earth's surface. It is not exactly the same, since $9.82\ \text{m s}^{-2}$ is an average for the whole Earth, the radius of the Earth being maximum at the equator.



Exercise

1 The mass of the Moon is 7.35×10^{22} kg and the radius 1.74×10^{3} km. What is the acceleration due to gravity on the Moon's surface?

How often does the Earth go around the Sun?

Applying Newton's universal law, we find that the force experienced by the Earth is given by:

$$F = G\frac{m_1 m_2}{r^2}$$

where

m_1 = mass of the Sun = 1.99×10^{30} kg

m_2 = mass of the Earth = 5.97×10^{24} kg

r = distance between the Sun and Earth = 1.49×10^{11} m

So $F = 3.56 \times 10^{22}$ N

W To build your own solar system with the 'solar system' simulation from PhET, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 6.1.

The planets orbit the Sun.

We know that the Earth travels in an elliptical orbit around the Sun, but we can take this to be a circular orbit for the purposes of this calculation. From our knowledge of circular motion we know that the force acting on the Earth towards the centre of the circle is the centripetal force given by the equation $F = \frac{mv^2}{r}$

So the velocity $v = \sqrt{\frac{Fr}{m}}$

$= 29\,846\ \text{m s}^{-1}$

The circumference of the orbit $= 2\pi r = 9.38 \times 10^{11}$ m

Time taken for 1 orbit $= \frac{9.38 \times 10^{11}}{29\,846}$

$= 3.14 \times 10^{7}$ s

This is equal to 1 year.

This agrees with observation. Newton's law has therefore predicted two correct results.

Here the law is used to make predictions that can be tested by experiment.

Gravitational field

The fact that both the apple and the Earth experience a force without being in contact makes gravity a bit different from the other forces we have come across. To model this situation, we introduce the idea of a *field*. A field is simply a region of space where something is to be found. A potato field, for example, is a region where you find potatoes. A gravitational field is a region where you find gravity. More precisely, gravitational field is defined as a region of space where a mass experiences a force because of its mass.

So there is a gravitational field in your classroom since masses experience a force in it.

Field strength on the Earth's surface:
Substituting M = mass of the Earth $= 5.97 \times 10^{24}$ kg
r = radius of the Earth = 6367 km
gives $g = Gm_1M/r^2$
$= 9.82\ \text{N kg}^{-1}$
This is the same as the acceleration due to gravity, which is what you might expect, since Newton's 2nd law says $a = F/m$.

Gravitational field strength (*g*)

This gives a measure of how much force a body will experience in the field. It is defined as the force per unit mass experienced by a small test mass placed in the field.

So if a test mass, m, experiences a force F at some point in space, then the field strength, g, at that point is given by $g = \frac{F}{m}$.

g is measured in N kg^{-1}, and is a vector quantity.

Note: The reason a small test mass is used is because a big mass might change the field that you are trying to measure.

Gravitational field around a spherical object

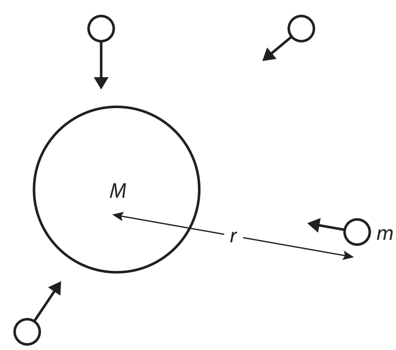


Figure 6.4 The region surrounding M is a gravitational field since all the test masses experience a force.

The force experienced by the mass, m is given by;

$$F = G\frac{Mm}{r^2}$$

So the field strength at this point in space, $g = \frac{F}{m}$

So

$$g = G\frac{M}{r^2}$$

Exercises

2 The mass of Jupiter is 1.89×10^{27} kg and the radius 71 492 km.
What is the gravitational field strength on the surface of Jupiter?

3 What is the gravitational field strength at a distance of 1000 km from the surface of the Earth?

Field lines

Field lines are drawn in the direction that a mass would accelerate if placed in the field – they are used to help us visualize the field.

The field lines for a spherical mass are shown in Figure 6.5.

The arrows give the direction of the field.

The field strength (g) is given by the density of the lines.

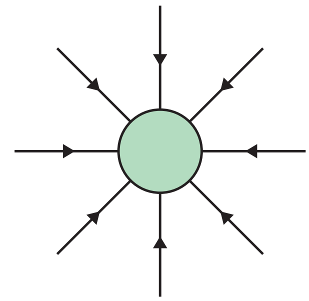


Figure 6.5 Field lines for a sphere of mass.

Gravitational field close to the Earth

When we are doing experiments close to the Earth, in the classroom for example, we assume that the gravitational field is uniform. This means that wherever you put a mass in the classroom it is always pulled downwards with the same force. We say that the field is *uniform*.

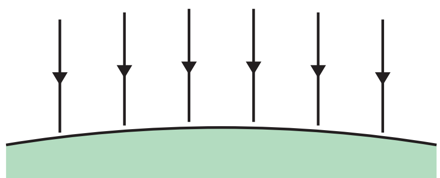


Figure 6.6 Regularly spaced parallel field lines imply that the field is uniform.

Addition of field

Since field strength is a vector, when we add field strengths caused by several bodies, we must remember to add them vectorially.

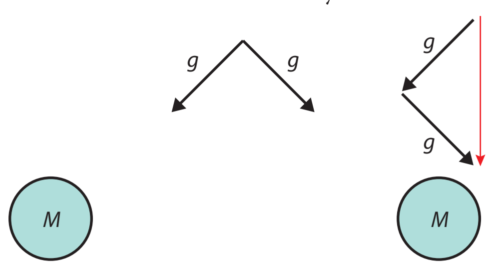


Figure 6.7 Vector addition of field strength.

In this example, the angle between the vectors is 90°. This means that we can use Pythagoras to find the resultant.

$$g = \sqrt{g_1^{\,2} + g_2^{\,2}}$$

Worked example

Calculate the gravitational field strength at points A and B in Figure 6.8.

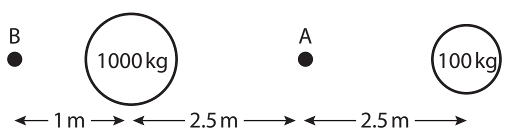


Figure 6.8

Solution

The gravitational field strength at A is equal to the sum of the field due to the two masses.

Field strength due to large mass $= G \times 1000/2.5^2 = 1.07 \times 10^{-8}\ \text{N kg}^{-1}$

Field strength due to small mass $= G \times 100/2.5^2 = 1.07 \times 10^{-9}\ \text{N kg}^{-1}$

Field strength $= 1.07 \times 10^{-8} - 1.07 \times 10^{-9}$

$= 9.63 \times 10^{-9}\ \text{N kg}^{-1}$

● **Examiner's hint:** Since field strength g is a vector, the resultant field strength equals the vector sum.

Exercises

4 Calculate the gravitational field strength at point B.

5 Calculate the gravitational field strength at A if the big mass were changed for a 100 kg mass.

6.2 Gravitational potential

Assessment statements

9.2.1 Define gravitational potential and gravitational potential energy.

9.2.2 State and apply the expression for gravitational potential due to a point mass.

9.2.3 State and apply the formula relating gravitational field strength to gravitational potential gradient.

9.2.4 Determine the potential due to one or more point masses.

9.2.5 Describe and sketch the pattern of equipotential surfaces due to one and two point masses.

9.2.6 State the relation between equipotential surfaces and gravitational field lines.

Gravitational potential in a uniform field

As you lift a mass m from the ground, you do work. This increases the PE of the object. As PE $= mgh$, we know the PE gained by the mass depends partly on the size of the mass (m) and partly on where it is (gh). The 'where it is' part is called the 'gravitational potential (V)'. This is a useful quantity because, if we know it, we can calculate how much PE a given mass would have if placed there.

Rearranging the equation for PE we get $gh = \frac{\text{PE}}{m}$ so potential is the PE per unit mass and has units J kg^{-1}.

Contours
Close to the Earth, lines of equipotential join points that are the same height above the ground. These are the same as contours on a map.

In the simple example of masses in a room, the potential is proportional to height, so a mass m placed at the same height in the room will have the same PE. By joining all positions of the same potential we get a line of equal potential, and these are useful for visualizing the changes in PE as an object moves around the room.

Worked examples

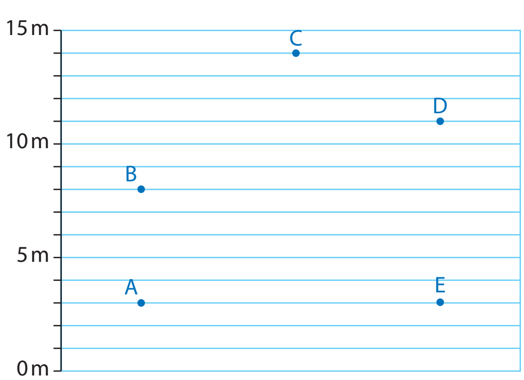


Figure 6.9

Referring to Figure 6.9.

1 What is the potential at A?

2 If a body is moved from A to B what is the change in potential?

3 How much work is done moving a 2 kg mass from A to B?

Solutions

1 $V_A = gh$ so potential at A $= 10 \times 3 = 30\text{ J kg}^{-1}$

2 $V_A = 30\text{ J kg}^{-1}$
$V_B = 80\text{ J kg}^{-1}$
Change in potential $= 80 - 30 = 50\text{ J kg}^{-1}$

3 The work done moving from A to B is equal to the change in potential $\times$ mass $= 50 \times 2 = 100\text{ J}$

Exercises

6 What is the difference in potential between C and D?

7 How much work would be done moving a 3 kg mass from D to C?

8 What is the PE of a 3 kg mass placed at B?

9 What is the potential difference between A and E?

10 How much work would be done taking a 2 kg mass from A to E?

Equipotentials and field lines

If we draw the field lines in our 15 m room they will look like Figure 6.10. The field is uniform so they are parallel and equally spaced. If you were to move upwards along a field line (A–B), you would have to do work and therefore your PE would increase. On the other hand, if you travelled perpendicular to the field lines (A–E), no work would be done, in which case you must be travelling along a line of equipotential. For this reason, field lines and equipotentials are perpendicular.

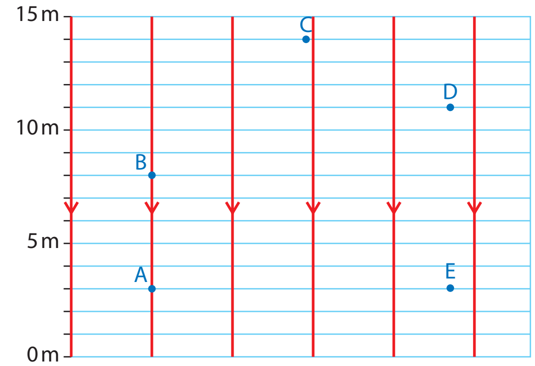


Figure 6.10 Equipotentials and field lines.

The amount of work done as you move up is equal to the change in potential × mass

Work = ΔVm

But the work done is also equal to

force × distance = $mg\Delta h$

So $\Delta Vm = mg\Delta h$

Rearranging gives

$$\frac{\Delta V}{\Delta h} = g$$

or the potential gradient = the field strength.

So lines of equipotential that are close together represent a strong field.

This is similar to the situation with contours as shown in Figure 6.11. Contours that are close together mean that the gradient is steep and where the gradient is steep, there will be a large force pulling you down the slope.

Figure 6.11 Close contours mean a steep mountain.

In this section we have been dealing with the simplified situation. Firstly we have only been dealing with bodies close to the Earth, where the field is uniform, and secondly we have been assuming that the ground is a position of zero potential. A more general situation would be to consider large distances away from a sphere of mass. This is rather more difficult but the principle is the same, as are the relationships between field lines and equipotentials.

Gravitational potential due to a massive sphere

The gravitational potential at point P is defined as:

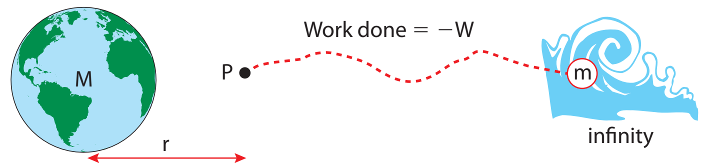
The work done per unit mass taking a small test mass from a position of zero potential to the point P.

In the previous example we took the Earth's surface to be zero but a better choice would be somewhere where the mass isn't affected by the field at all. Since $g = \frac{GM}{r^2}$ the only place completely out of the field is at an infinite distance from the mass – so let's start there.

Infinity
We can't really take a mass from infinity and bring it to the point in question, but we can calculate how much work would be required if we did. Is it OK to calculate something we can never do?

Figure 6.12 represents the journey from infinity to point P, a distance r from a large mass M. The work done making this journey $= -W$ so the potential $V = \frac{-W}{m}$

Figure 6.12 The journey from infinity to point *P*.

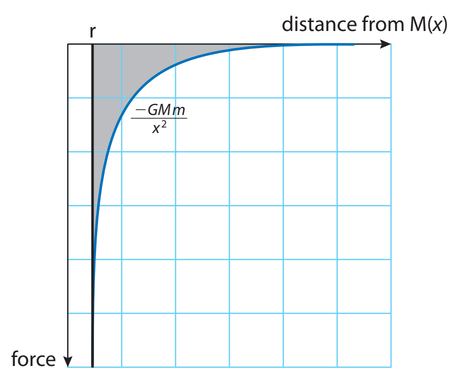


The negative sign is because if you were taking mass m from infinity to P you wouldn't have to pull it, it would pull you. The direction of the force that you would exert is in the opposite direction to the way it is moving, so work done is negative.

Calculating the work done

There are two problems when you try to calculate the work done from infinity to P; firstly the distance is infinite (obviously) and secondly the force gets bigger as you get closer. To solve this problem, we use the area under the force–distance graph (remember the work done stretching a spring?). From Newton's universal law of gravitation we know that the force varies according to the equation: $F = \frac{GMm}{r^2}$ so the graph will be as shown in Figure 6.13.

Figure 6.13 Graph of force against distance as the test mass is moved towards *M*.



Integration

The integral mentioned here is

$$V = \int_{\infty}^{r} \frac{GM}{x^2}\,dx$$

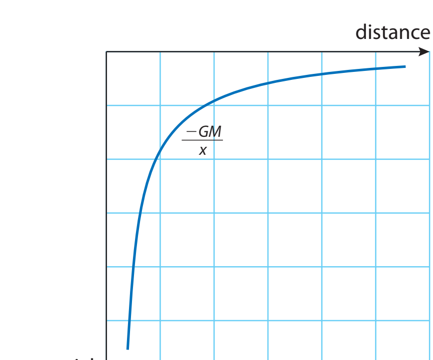
The area under this graph can be found by integrating the function $-\frac{GMm}{x^2}$ from infinity to r (you'll do this in maths). This gives the result:

$$W = -\frac{GMm}{r}$$

So the potential, $V = \frac{W}{m} = -\frac{GM}{r}$

The graph of potential against distance is drawn in Figure 6.14. The gradient of this line gives the field strength, but notice that the gradient is positive and the field strength negative so we get the formula $g = -\frac{\Delta V}{\Delta x}$

Figure 6.14 Graph of potential against distance.



Equipotentials and potential wells

If we draw the lines of equipotential for the field around a sphere, we get concentric circles, as in Figure 6.15. In 3D these would form spheres, in which case they would be called *equipotential surfaces* rather the lines of equipotential.

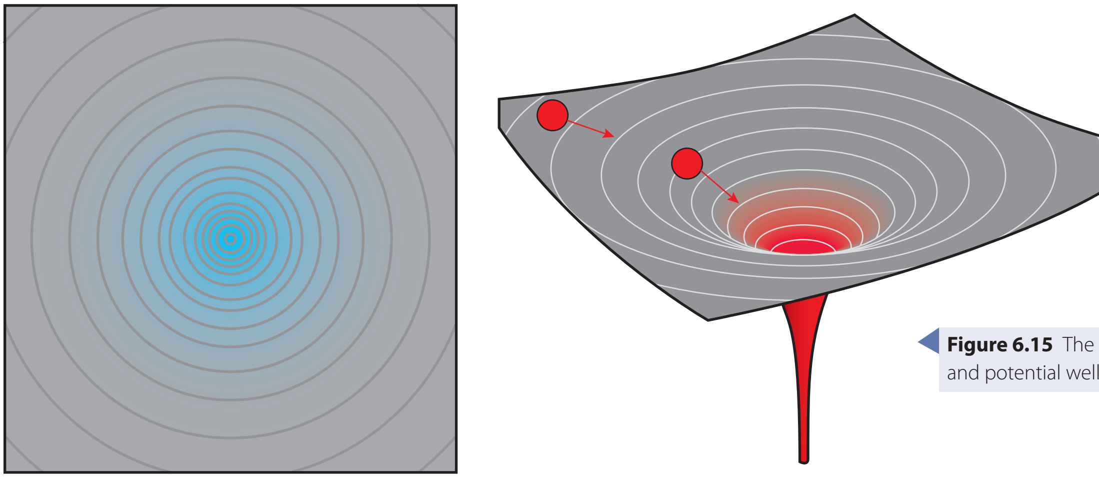


Figure 6.15 The lines of equipotential and potential well for a sphere.

An alternative way of representing this field is to draw the hole or well that these contours represent. This is a very useful visualization, since it not only represents the change in potential but by looking at the gradient, we can also see where the force is biggest. If you imagine a ball rolling into this well you can visualize the field.

Relationship between field lines and potential

If we draw the field lines and the potential as in Figure 6.16, we see that as before they are perpendicular. We can also see that the lines of equipotential are closest together where the field is strongest (where the field lines are most dense). This agrees with our earlier finding that $g = \dfrac{-\Delta V}{\Delta x}$

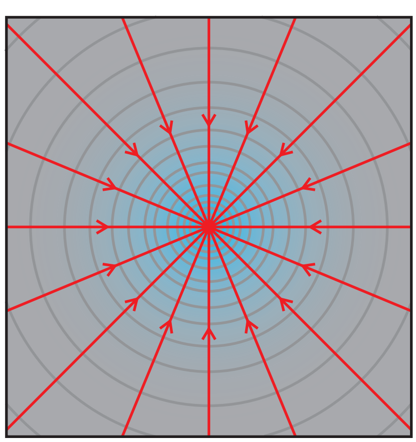
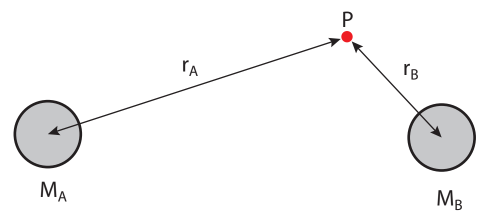


Figure 6.16 Equipotentials and field lines.

Addition of potential

Potential is a scalar quantity, so adding potentials is just a matter of adding the magnitudes. If we take the example shown in Figure 6.17, to find the potential at point P we calculate the potential due to A and B then add them together.

Figure 6.17 Two masses.

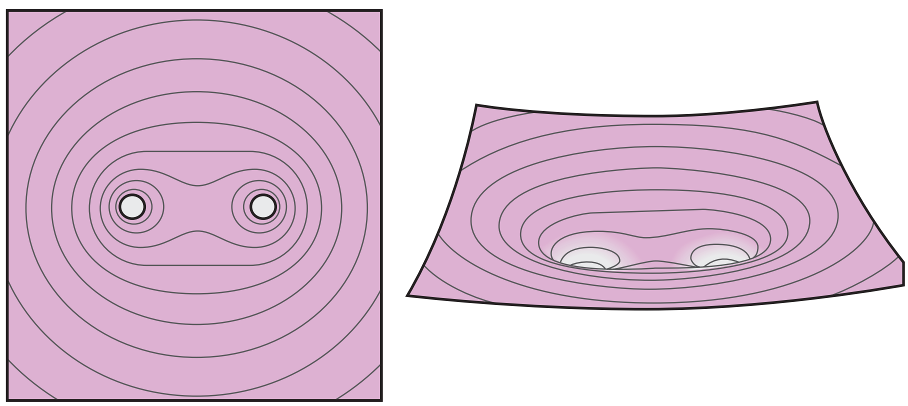


● **Hint:** When you add field strengths you have to add them vectorially, making triangles using Pythagoras etc. Adding potential is much simpler because it's a scalar.

The total potential at P $= -\frac{GM_A}{r_A} + -\frac{GM_B}{r_B}$

The lines of equipotential for this example are shown in Figure 6.18.

Figure 6.18 Equipotentials and potential wells for two equal masses. If you look at the potential well, you can imagine a ball could sit on the hump between the two holes. This is where the field strength is zero.



Exercise

11 The Moon has a mass of 7.4×10^{22} kg and the Earth has mass of 6.0×10^{24} kg. The average distance between the Earth and the Moon is 3.8×10^5 km. If you travel directly between the Earth and the Moon in a rocket of mass 2000 kg

(a) calculate the gravitational potential when you are 1.0×10^4 km from the Moon

(b) calculate the rocket's PE at the point in part **(a)**

(c) draw a sketch graph showing the change in potential

(d) mark the point where the gravitational field strength is zero.

6.3 Escape speed

Assessment statements

9.2.7 Explain the concept of escape speed from a planet.

9.2.8 Derive an expression for the escape speed of an object from the surface of a planet.

9.2.9 Solve problems involving gravitational potential energy and gravitational potential.

If a body is thrown straight up, its KE decreases as it rises. If we ignore air resistance, this KE is converted into PE. When it gets to the top, the final PE will equal the initial KE, so $\frac{1}{2}mv^2 = mgh$.

If we throw a body up really fast, it might get so high that the gravitational field strength would start to decrease. In this case, we would have to use the formula for the PE around a sphere.

$$\text{PE} = -\frac{GMm}{r}$$

So when it gets to its furthest point as shown in Figure 6.19

$$\text{loss of KE} = \text{gain in PE}$$

$$\tfrac{1}{2}mv^2 - 0 = -\frac{GMm}{R_2} - -\frac{GMm}{R_E}$$

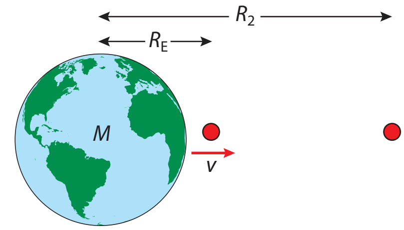


Figure 6.19 A mass *m* thrown away from the Earth.

If we throw the ball fast enough, it will never come back. This means that it has reached a place where it is no longer attracted back to the Earth, infinity. Of course it can't actually reach infinity but we can substitute $R_2 = \infty$ into our equation to find out how fast that would be.

$$\tfrac{1}{2}mv^2 = -\frac{GMm}{\infty} - -\frac{GMm}{R_E}$$

$$\tfrac{1}{2}mv^2 = \frac{GMm}{R_E}$$

Rearranging gives: $$v_{escape} = \sqrt{\frac{2GM}{R_E}}$$

If we calculate this for the Earth it is about $11\ \text{km s}^{-1}$.

Air resistance

If you threw something up with a velocity of $11\ \text{km s}^{-1}$ it would be rapidly slowed by air resistance. The work done against this force would be converted to thermal energy causing the body to vaporize. Rockets leaving the Earth do not have to travel anywhere near this fast, as they are not thrown upwards, but have a rocket engine that provides a continual force.

Why the Earth has an atmosphere but the Moon doesn't

The average velocity of an air molecule at the surface of the Earth is about $500\ \text{m s}^{-1}$. This is much less than the velocity needed to escape from the Earth, and for that reason the atmosphere doesn't escape.

The escape velocity on the Moon is $2.4\ \text{km s}^{-1}$ so you might expect the Moon to have an atmosphere. However, $500\ \text{m s}^{-1}$ is the *average* speed; a lot of the molecules would be travelling faster than this leading to a significant number escaping, and over time all would escape.

Black holes

A star is a big ball of gas held together by the gravitational force. The reason this force doesn't cause the star to collapse is that the particles are continuously given KE from the nuclear reactions taking place (fusion). As time progresses, the nuclear fuel gets used up, so the star starts to collapse. As this happens, the escape velocity increases until it is bigger than the speed of light, at this point not even light can escape and the star has become a black hole.

Exercises

12 The mass of the Moon is 7.4×10^{22} kg and its radius is 1738 km. Show that its escape speed is $2.4\ \text{km s}^{-1}$.

13 Why doesn't the Earth's atmosphere contain hydrogen?

14 The mass of the Sun is 2.0×10^{30} kg. Calculate how small its radius would have to be for it to become a black hole.

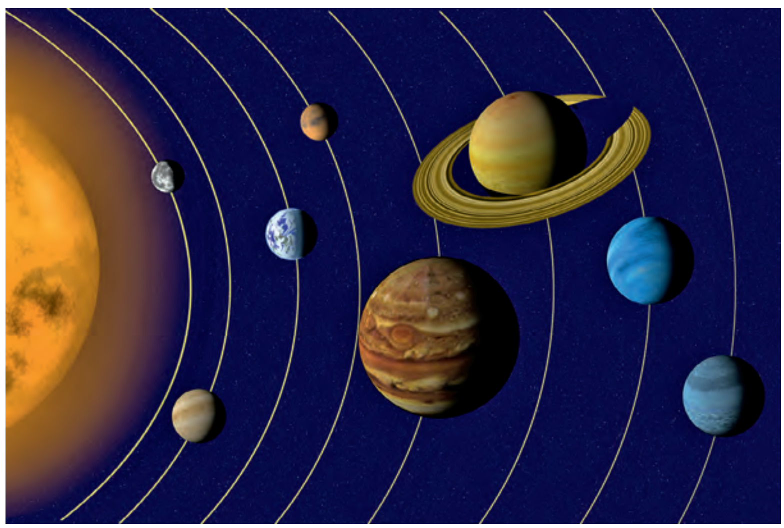
15 When travelling away from the Earth, a rocket runs out of fuel at a distance of 1.0×10^5 km. How fast would the rocket have to be travelling for it to escape from the Earth? (Mass of the Earth $= 6.0 \times 10^{24}$ kg, radius $= 6400$ km.)

How can light be slowed down by the effect of gravity, when according to Newton's law, it has no mass, therefore isn't affected by gravity? This can't be explained by Newton's theories, but Einstein solved the problem with his general theory of relativity.

6.4 Orbital motion

Assessment statements

9.4.1 State that gravitation provides the centripetal force for circular orbital motion.
9.4.2 Derive Kepler's third law.
9.4.3 Derive expressions for the kinetic energy, potential energy and total energy of an orbiting satellite.
9.4.4 Sketch graphs showing the variation with orbital radius of the kinetic energy, gravitational potential energy and total energy of a satellite.
9.4.5 Discuss the concept of weightlessness in orbital motion, in free fall and in deep space.
9.4.6 Solve problems involving orbital motion.



An artist's impression of the solar system.

The solar system

The solar system consists of the Sun at the centre surrounded by eight orbiting planets. The shape of the orbits is actually slightly elliptical but to make things simpler, we will assume them to be circular. We know that for a body to travel in a circle, there must be an unbalanced force (called the centripetal force, $m\omega^2r$) acting towards the centre. The force that holds the planets in orbit around the Sun is the gravitational force $\frac{GMm}{r^2}$. Equating these two expressions gives us an equation for orbital motion.

$$m\omega^2 r = \frac{GMm}{r^2} \qquad (1)$$

Now ω is the angular speed of the planet, that is the angle swept out by a radius per unit time. If the time taken for one revolution (2π radians) is T then $\omega = \frac{2\pi}{T}$.

Substituting into equation (1) gives

$$m\left(\frac{2\pi}{T}\right)^2 r = \frac{GMm}{r^2}$$

Rearranging gives: $$\frac{T^2}{r^3} = \frac{4\pi^2}{GM}$$

where M is the mass of the Sun.

So for planets orbiting the Sun, $\frac{T^2}{r^3}$ is a constant, or T^2 is proportional to r^3.

This is Kepler's third law.

Kepler thought of his law before Newton was born, so couldn't have derived the equation in the way we have here. He came up with the law by manipulating the data that had been gathered from many years of measurement, realizing that if you square the time period and divide by the radius cubed you always get the same number.

From this we can deduce that the planet closest to the Sun (Mercury) has a shorter time period than the planet furthest away. This is supported by measurement:

Time period of Mercury = 0.24 years.

Time period of Neptune = 165 years.

Exercise

16 Use a database to make a table of the values of time period and radius for all the planets. Plot a graph to show that T^2 is proportional to r^3.

To find an example of a database, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 6.2.

Energy of an orbiting body

As planets orbit the Sun they have KE due to their movement and PE due to their position. We know that their PE is given by the equation:

$$\text{PE} = -\frac{GMm}{r}$$

and $$\text{KE} = \frac{1}{2}mv^2$$

We also know that if we approximate the orbits to be circular then equating the centripetal force with gravity gives:

$$\frac{GMm}{r^2} = \frac{mv^2}{r}$$

Rearranging and multiplying by $\frac{1}{2}$ gives $$\frac{1}{2}mv^2 = \frac{GMm}{2r}$$

$$\text{KE} = \frac{GMm}{2r}$$

The total energy $= \text{PE} + \text{KE} = -\frac{GMm}{r} + \frac{GMm}{2r}$

Total energy $= -\frac{GMm}{2r}$

● **Hint:** There are two versions of the equation for centripetal force

Speed version:

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

Angular speed version:

$F = m\omega^2 r$

Earth satellites

The equations we have derived for the orbits of the planets also apply to the satellites that man has put into orbits around the Earth. This means that the satellites closer to the Earth have a time period much shorter than the distant ones. For example, a low orbit spy satellite could orbit the Earth once every two hours and a much higher TV satellite orbits only once a day.

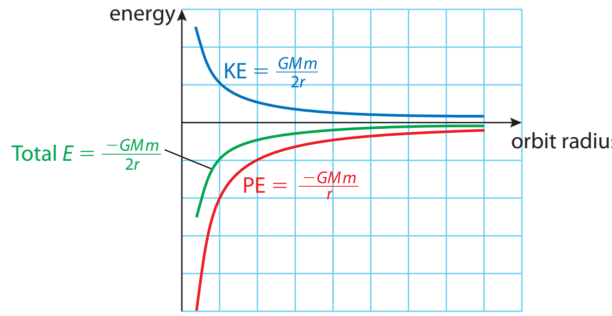
The total energy of an orbiting satellite $= -\frac{GMm}{2r}$ so the energy of a high satellite (big r) is less negative and hence bigger than a low orbit. To move from a low orbit into a high one therefore requires energy to be added (work done).

Imagine you are in a spaceship orbiting the Earth in a low orbit. To move into a higher orbit you would have to use your rocket motor to increase your energy. If you kept doing this you could move from orbit to orbit, getting further and further from the Earth. The energy of the spaceship in each orbit can be displayed as a graph as in Figure 6.20 (overleaf).

From the graph we can see that low satellites have greater KE but less total energy than distant satellites, so although the distant ones move with slower speed, we have to do work to increase the orbital radius. Going the other way, to move from a distant orbit to a close orbit, the spaceship needs to lose energy. Satellites in low

Earth orbit are not completely out of the atmosphere, so lose energy due to air resistance. As they lose energy they spiral in towards the Earth.

Figure 6.20 Graph of KE, PE and total energy for a satellite with different orbital radius.



The physicist Stephen Hawking experiencing weightlessness in a free falling aeroplane.

Weightlessness

The only place you can be truly without weight is a place where there is no gravitational field; this is at infinity or a place where the gravitational fields of all the bodies in the universe cancel out. If you are a long way from everything, somewhere in the middle of the universe, then you could say that you are pretty much weightless.

To understand how it feels to be weightless, we first need to think what it is that makes us feel weight. As we stand in a room we can't feel the Earth pulling our centre downwards but we can feel the ground pushing our feet up. This is the normal force that must be present to balance our weight. If it were not there we would be accelerating downwards. Another thing that makes us notice that we are in a gravitational field is what happens to things we drop; it is gravity that pulls them down. Without gravity they would float in mid-air. So if we were in a place where there was no gravitational field then the floor would not press on our feet and things we drop would not fall. It would feel exactly the same if we were in a room that was falling freely as in Figure 6.21. If we accelerate down along with the room then the only force acting on us is our weight; there is no normal force between the floor and our feet. If we drop something it falls with us. From outside the room we can see that the room is in a gravitational field falling freely but inside the room it feels like someone has turned off gravity (not for long though). An alternative and rather longer lasting way of feeling weightless is to orbit the Earth inside a space station. Since the space station and everything inside it is accelerating towards the Earth, it will feel exactly like the room in Figure 6.21, except you won't hit the ground.

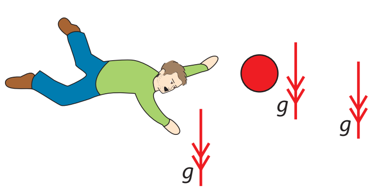


Figure 6.21 As the room, the man and the ball accelerate downwards, the man will feel weightless.

Exercises

17 So that they can stay above the same point on the Earth, TV satellites have a time period equal to one day. Calculate the radius of their orbit.

18 A spy satellite orbits 400 km above the Earth. If the radius of the Earth is 6400 km, what is the time period of the orbit?

19 If the satellite in question 18 has a mass of 2000 kg, calculate its

(a) KE

(b) PE

(c) total energy.

6.5 Electric force and field

Assessment statements

6.2.1 State that there are two types of electric charge.
6.2.2 State and apply the law of conservation of charge.
6.2.3 Describe and explain the difference in the electrical properties of conductors and insulators.
6.2.4 State Coulomb's law.
6.2.5 Define *electric field strength*.
6.2.6 Determine the electric field strength due to one or more point charges.
6.2.7 Draw the electric field patterns for different charge configurations.
6.2.8 Solve problems involving electric charges, forces and fields.

Electric force

So far we have dealt with many forces; for example, friction, tension, upthrust, normal force, air resistance and gravitational force. If we rub a balloon on a woolen pullover, we find that the balloon is attracted to the wool of the pullover – this cannot be explained in terms of any of the forces we have already considered, so we need to develop a new model to explain what is happening. First we need to investigate the effect.

This is another example of how models are used in physics.

Consider a balloon and a woollen pullover – if the balloon is rubbed on the pullover, we find that it is attracted to the pullover. However, if we rub two balloons on the pullover, the balloons repel each other.

Whatever is causing this effect must have two different types, since there are two different forces. We call this force the *electric force*.

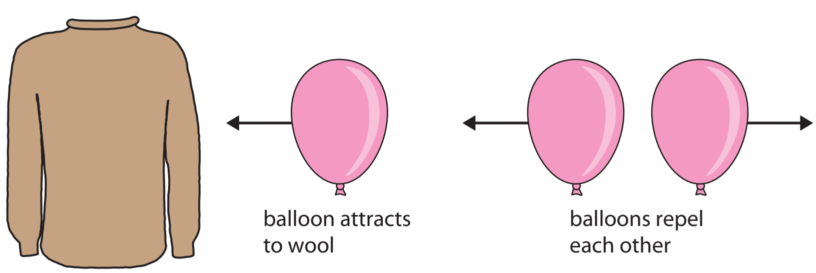


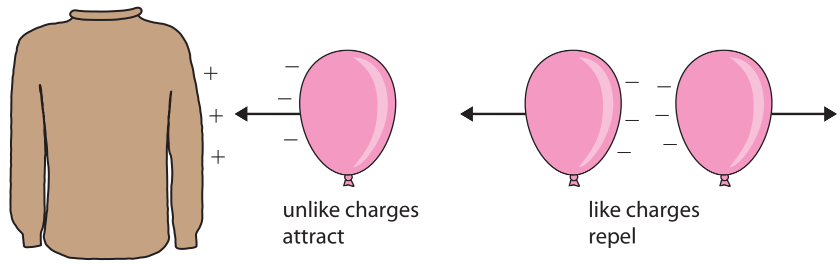
Figure 6.22 Balloons are attracted to the wool but repel each other.

Charge

The balloon and pullover must have some property that is causing this force. We call this property *charge*. There must be two types of charge, traditionally called positive (+ve) and negative(−ve). To explain what happens, we can say that, when rubbed, the balloon gains −ve charge and the pullover gains +ve charge. If like charges repel and unlike charges attract, then we can explain why the balloons repel and the balloon and pullover attract.

You can try this with real balloons or, to use the simulation 'Balloons and static electricity', visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 6.3.

Figure 6.23 The force is due to charges.



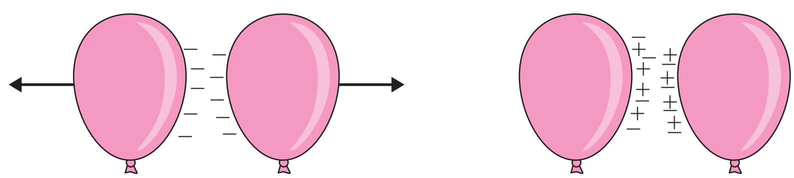
Here +ve and −ve numbers are used to represent something that they were not designed to represent.

The unit of charge is the coulomb (C).

Conservation of charge

If we experiment further, we find that if we rub the balloon more, then the force between the balloons is greater. We also find that if we add +ve charge to an equal −ve charge, the charges cancel.

Figure 6.24 Charges cancel each other out.



We can add and take away charge but we cannot destroy it.

The law of conservation of charge states that charge can neither be created nor destroyed.

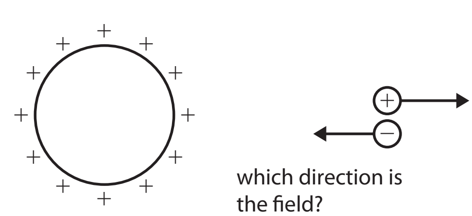
Electric field

We can see that there are certain similarities with the electric force and gravitational force; they both act without the bodies touching each other. We used the concept of a field to model gravitation and we can use the same idea here.

Electric field is defined as a region of space where a charged object experiences a force due to its charge.

Field lines

Figure 6.25 Field is in the direction in which a +ve charge would accelerate.



Field lines can be used to show the direction and strength of the field. However, because there are two types of charge, the direction of the force could be one of two possibilities.

It has been decided that we should take the direction of the field to be the direction that a small +ve charge would accelerate if placed in the field. So we will always consider what would happen if +ve charges are moved around in the field. The field lines will therefore be as shown in Figures 6.26 to 6.28.

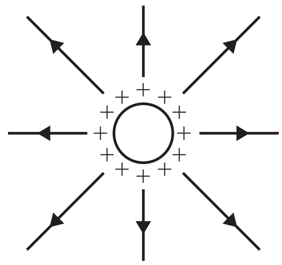


Figure 6.26 Field lines close to a sphere of charge.

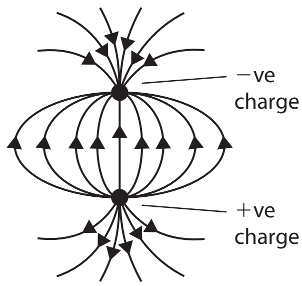


Figure 6.27 Field due to a dipole.

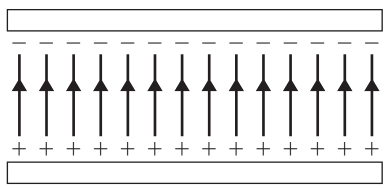


Figure 6.28 A uniform field.

Coulomb's law

In a gravitational field, the force between masses is given by Newton's law, and the equivalent for an electric field is Coulomb's law.

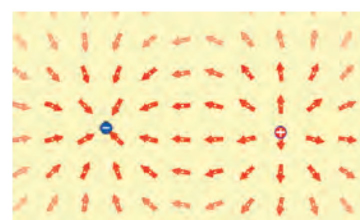
Coulomb's law states that the force experienced by two point charges is directly proportional to the product of their charge and inversely proportional to the square of their separation.

The force experienced by two point charges Q_1 and Q_2 separated by a distance r in a vacuum is given by the formula

$$F = k\frac{Q_1 Q_2}{r^2}$$

The constant of proportionality $k = 9 \times 10^9\ \mathrm{Nm^2\,C^{-2}}$

Note: Similarly to gravitational fields, Coulomb's law also applies to spheres of charge, the separation being the distance between the centres of the spheres.



W In the PhET simulation 'Charges and fields' you can investigate the force experienced by a small charge as it is moved around an electric field. To try this, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 6.4.

Electric field strength (*E*)

The electric field strength is a measure of the force that a +ve charge will experience if placed at a point in the field. It is defined as the force per unit charge experienced by a small +ve test charge placed in the field.

So if a small +ve charge q experiences a force F in the field, then the field strength at that point is given by $E = \frac{F}{q}$. The unit of field strength is $\mathrm{NC^{-1}}$, and it is a vector quantity.

Worked examples

A 5 μC point charge is placed 20 cm from a 10 μC point charge.

1 Calculate the force experienced by the 5 μC charge.

2 What is the force on the 10 μC charge?

3 What is the field strength 20 cm from the 10 μC charge?

Solutions

1 Using the equation $F = \frac{kQ_1Q_2}{r^2}$

$Q_1 = 5 \times 10^{-6}$ C, $Q_2 = 10 \times 10^{-6}$ C and $r = 0.20$ m

$$F = \frac{9 \times 10^9 \times 5 \times 10^{-6} \times 10 \times 10^{-6}}{0.20^2}\ \text{N}$$
$$= 11.25\ \text{N}$$

2 According to Newton's third law, the force on the 10 μC charge is the same as the 5 μC.

3 Force per unit charge $= \frac{11.25}{5 \times 10^{-6}}$

$E = 2.25 \times 10^6\ \text{N C}^{-1}$

● **Examiner's hint:** Field strength is defined as the force per unit charge so if the force on a 5 μC charge is 11.25 N, the field strength E is equal to 11.25 N divided by 5 μC.

Exercises

20 If the charge on a 10 cm radius metal sphere is 2 μC, calculate
(a) the field strength on the surface of the sphere
(b) the field strength 10 cm from the surface of the sphere
(c) the force experienced by a 0.1 μC charge placed 10 cm from the surface of the sphere.

21 A small sphere of mass 0.01 kg and charge 0.2 μC is placed at a point in an electric field where the field strength is 0.5 N C^{-1}.
(a) What force will the small sphere experience?
(b) If no other forces act, what is the acceleration of the sphere?

● **Examiner's hint:** When solving field problems you always assume one of the charges is in the field of the other. E.g in Example 1 the 5 μC charge is in the field of the 10 μC charge. Don't worry about the fact that the 5 μC charge also creates a field – that's not the field you are interested in.

Electric field strength in a uniform field

A uniform field can be created between two parallel plates of equal and opposite charge as shown in Figure 6.29. The field lines are parallel and equally spaced. If a test charge is placed in different positions between the plates, it experiences the same force.

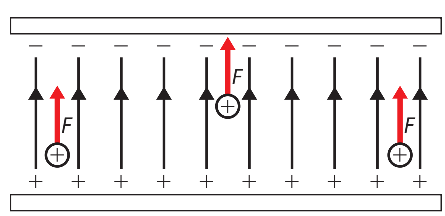


Figure 6.29 In a uniform field the force on a charge q is the same everywhere.

So if a test charge q is placed in the field above, then $E = \frac{F}{q}$ everywhere between the two charged plates.

Worked example

If a charge of 4 μC is placed in a uniform field of field strength 2 N C^{-1} what force will it experience?

Solution

$F = EQ$ Rearranging the formula $E = \frac{F}{Q}$

$= 2 \times 4 \times 10^{-6}$ N

$= 8$ μN

Electric field strength close to a sphere of charge

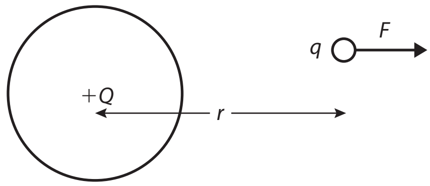


Figure 6.30

From definition: $E = \frac{F}{q}$

From Coulomb's law: $F = k\frac{Qq}{r^2}$

Substituting: $E = k\frac{Q}{r^2}$

Addition of field strength

Field strength is a vector, so when the field from two negatively charged bodies act at a point, the field strengths must be added vectorially. In Figure 6.31, the resultant field at two points A and B is calculated. At A the fields act in the same line but at B a triangle must be drawn to find the resultant.

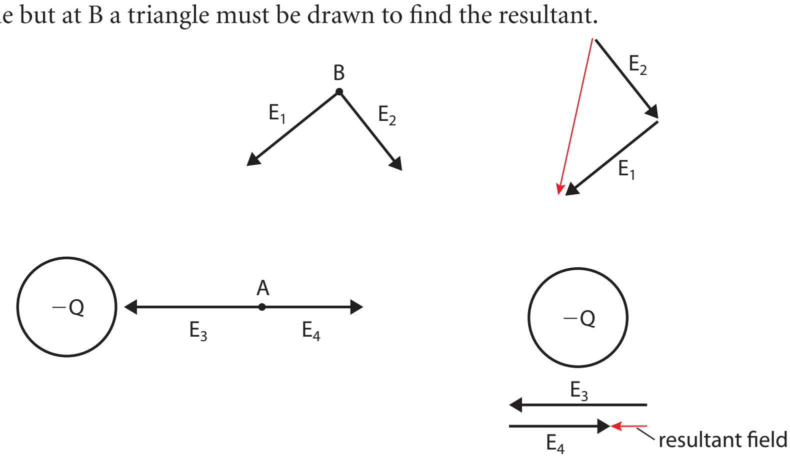
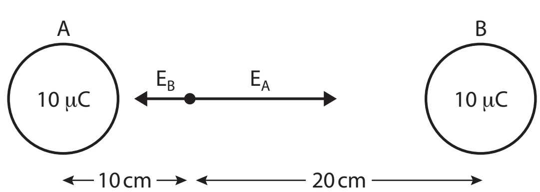


Figure 6.31 Since both charges are negative, the field strength is directed towards the charges. Since they are at right angles to each other, Pythagoras can be used to sum these vectors.

E_3 is the field due to the charge on the left, E_4 is due to the charge on the right. E_3 is bigger than E_4 because the charge on the left is closer.

Worked example



Two +10 μC charges are separated by 30 cm. What is the field strength between the charges 10 cm from A?

Solution

Field strength due to A, $E_A = \frac{9 \times 10^9 \times 10 \times 10^{-6}}{0.1^2}$

$= 9 \times 10^6\,\mathrm{N\,C^{-1}}$

$E_B = \frac{9 \times 10^9 \times 10 \times 10^{-6}}{0.2^2}$

$= 2.25 \times 10^6\,\mathrm{N\,C^{-1}}$

Resultant field strength $= (9 - 2.25) \times 10^6\,\mathrm{N\,C^{-1}}$

$= 6.75 \times 10^6\,\mathrm{N\,C^{-1}}$ (to the left)

6.6 Electrical potential

Assessment statements

9.3.1 Define electric potential and electric potential energy.
9.3.2 State and apply the expression for electric potential due to a point charge.
9.3.3 State and apply the formula relating electric field strength to electric potential gradient.
9.3.4 Determine the potential due to one or more point charges.
9.3.5 Describe and sketch the pattern of equipotential surfaces due to one and two point charges.
9.3.6 State the relation between equipotential surfaces and electric field lines.
9.3.7 Solve problems involving electric potential energy and electric potential.

The concept of electric potential is very similar to that of gravitational potential; it gives us information about the amount of energy associated with different points in a field. We have already defined electric potential difference in relation to electrical circuits; it is the amount of electrical energy converted to heat when a unit charge flows through a resistor. In this section we will define the electric potential in more general terms.

Electric potential energy and potential

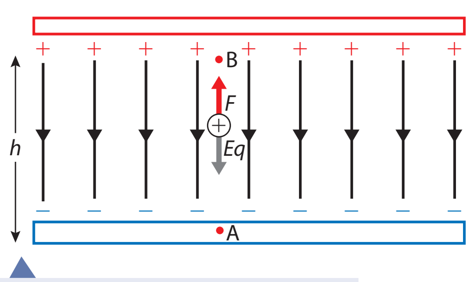


Figure 6.32

When we move a positive charge around in an electric field we have to do work on it. If we do work we must give it energy. This energy is not increasing the KE of the particle so must be increasing its PE, and so we call this *electric potential energy*. Let us first consider the uniform field shown in Figure 6.32. In order to move a charge $+q$ from A to B, we need to exert a force that is equal and opposite to the electric force, Eq. As we move the charge we do an amount of work equal to Eqh. We have therefore increased the electrical PE of the charge by the same amount, so PE $= Eqh$.

This is very similar to the room in Figure 6.9; the higher we lift the positive charge, the more PE it gets. In the same way we can define the potential of different points as being the quantity that defines how much energy a given charge would have if placed there.

Positive charge
Note that the potential is defined in terms of a positive charge.

Electric potential at a point is the amount of work per unit charge needed to take a small positive test charge from a place of zero potential to the point.

The unit of potential is J C^{-1} or volts.

Potential is a scalar quantity.

In this example if we define the zero in potential as the bottom plate, then the potential at B, $V_B = Eh$

Since the potential is proportional to h we can deduce that all points a distance h from the bottom plate will have the same potential. We can therefore draw lines of equipotential as we did in the gravitational field. Figure 6.33 shows an example with equipotentials.

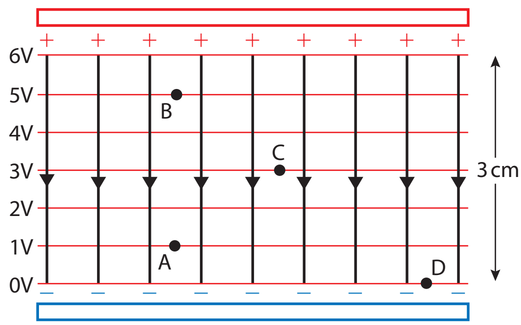


Figure 6.33

The electronvolt (eV)
This is a unit of energy used in atomic physics. 1 eV = energy gained by an electron accelerated through a p.d. of 1 V.

Exercises

Refer to Figure 6.33 for Questions 22–27.

22 What is the potential difference (p.d.) between A and C?

23 What is the p.d. between B and D?

24 If a charge of +3 C was placed at B, how much PE would it have?

25 If a charge of +2 C was moved from C to B, how much work would be done?

26 If a charge of −2 C moved from A to B, how much work would be done?

27 If a charge of +3 C was placed at B and released
(a) what would happen to it?
(b) how much KE would it gain when it reached A?

Contours
The lines of equipotential are again similar to contour lines but this time there is no real connection to gravity. Any hills and wells will be strictly imaginary.

Potential and field strength

In the example of a uniform field the change in potential ΔV when a charge is moved a distance Δh is given by

$$\Delta V = E\Delta h$$

Rearranging gives $E = \frac{\Delta V}{\Delta h}$, the field strength = the potential gradient. So in the example of Figure 6.33 the field strength is $\frac{6}{3 \times 10^{-2}} \frac{\text{V}}{\text{m}} = 200\ \text{N C}^{-1}$

Potential due to point charge

The uniform field is a rather special case; a more general example would be to consider the field due to a point charge. This would be particularly useful since all bodies are made of points; so if we know how to find the field due to one point we can find the field due to many points. In this way we can find the field caused by any charged object.

Consider a point P a distance r from a point charge Q as shown in Figure 6.34. The potential at P is defined as the work done per unit charge taking a small positive test charge from infinity to P.

Permittivity
The constant $k = 9 \times 10^9\ \text{Nm}^2\,\text{C}^{-2}$

This can also be expressed in terms of the permittivity of a vacuum, ϵ_0

$k = \frac{1}{4\pi\epsilon_0}$

$\epsilon_0 = 8.85 \times 10^{-12}\ \text{C}^2\,\text{N}^{-1}\,\text{m}^{-2}$

The permittivity is different for different media but we will only be concerned with fields in a vacuum.

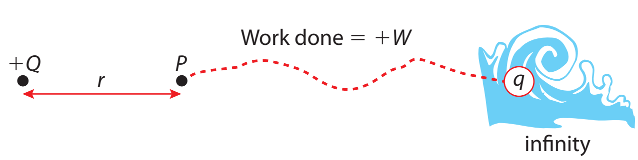


Figure 6.34 A positive charge is taken from infinity to point P.

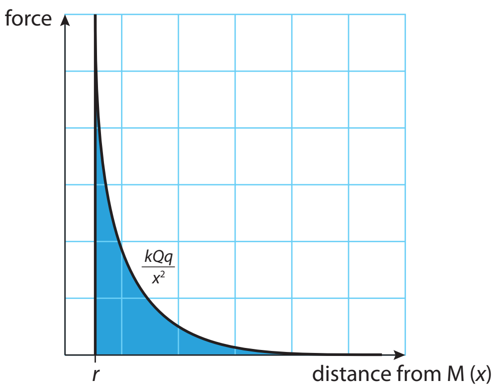


Figure 6.35 Graph of force against distance as charge $+q$ approaches $+Q$

Again we have used infinity as our zero of potential. To solve this we must find the area under the graph of force against distance, as we did in the gravity example.

Figure 6.35 shows the area that represents the work done. Notice that this time the force is positive as is the area under the graph. The area under this graph is given by the equation $\frac{kQq}{r}$ so the work done is given by:

$$W = \frac{kQq}{r}$$

The potential is the work done per unit charge so $V = \frac{kQ}{r}$.

The potential therefore varies as shown by the graph in Figure 6.36.

Potential gradient

The potential gradient is related to the field strength by the equation

$$\frac{dV}{dx} = -E$$

We can see that this is negative because the gradient of V against x is negative but the field strength is positive.

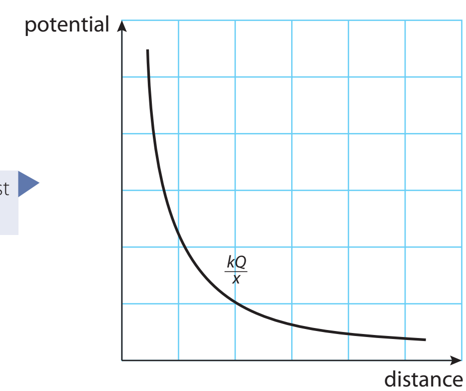


Figure 6.36 Graph of potential against distance for a positive charge

Equipotentials, wells and hills

If we draw lines of equipotential for a point charge we get concentric circles as shown in Figure 6.37. These look just the same as the gravitational equipotentials of Figure 6.15. However, we must remember that if the charge is positive, the potential increases as we get closer to it, rather than decreasing as in the case of gravitational field. These contours represent a hill not a well.

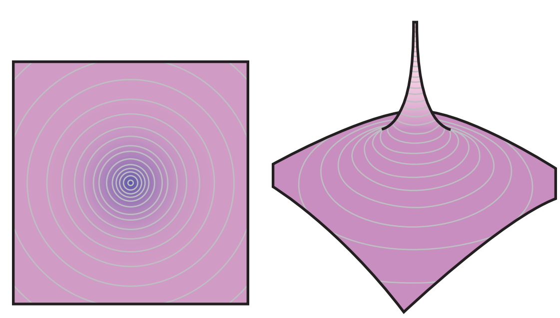


Figure 6.37 The equipotentials and potential hill of a positive point charge.

Exercises

28 Calculate the electric potential a distance 20 cm from the centre of a small sphere of charge $+50\ \mu C$.

29 Calculate the p.d. between the point in question 28 and a second point 40 cm from the centre of the sphere.

Addition of potential

Since potential is a scalar there is no direction to worry about when adding the potential from different bodies – simply add them together.

Example

At point P in Figure 6.38 the combined potential is given by:

$$V = k\frac{Q_1}{r_1} + k\frac{-Q_2}{r_2}$$

The potential of combinations of charging can be visualized by drawing lines of equipotentials. Figure 6.39 shows the equipotentials for a combination of a positive and negative charge (a dipole). This forms a hill and a well, and we can get a feeling for how a charge will behave in the field by imagining a ball rolling about on the surface shown.

● **Hint:** Zero potential is not zero field. If we look at the hill and well in Figure 6.39, we can see that there is a position of zero potential in between the two charges (where the potentials cancel). This is not, however, a position of zero field since the fields will both point towards the right in between the charges.

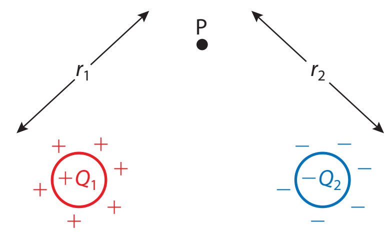


Figure 6.38

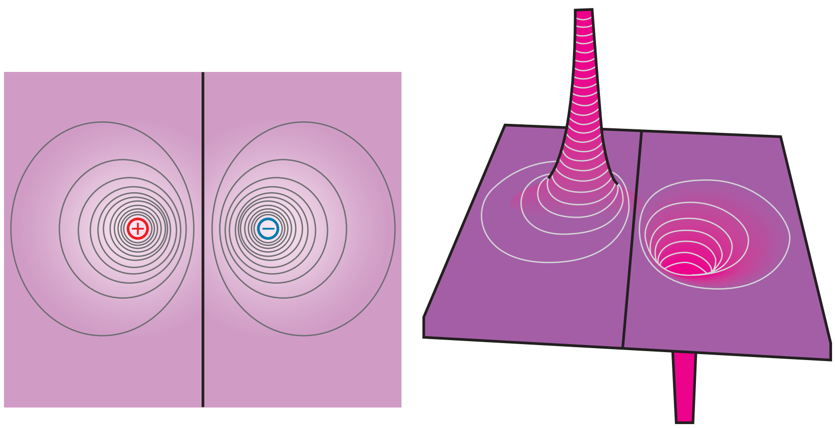


Figure 6.39 Equipotentials for a dipole.

Exercises

Refer to Figure 6.40 for questions 30–36.

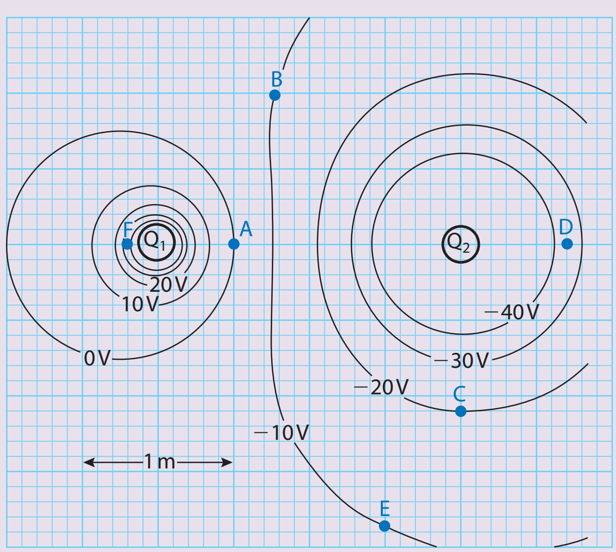


Figure 6.40 The lines of equipotential drawn every 10 V for two charges Q_1 and Q_2

30 (a) One of the charges is positive and the other is negative. Which is which?

(b) If a positive charge were placed at A, would it move, and if so, in which direction?

31 At which point A, B, C, D or F is the field strength greatest?

The electronvolt (eV)
The electronvolt is a unit of electrical PE often used in atomic and nuclear physics. 1eV is the amount of energy gained by an electron accelerated through a p.d. of 1 V.

32 What is the p.d. between the following pairs of points?
(a) A and C **(b)** C and E **(c)** B and E

33 How much work would be done taking a +2 C charge between the following points?
(a) C to A **(b)** E to C **(c)** B to E

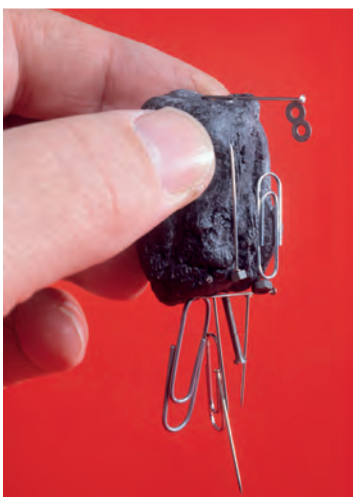
34 Using the scale on the diagram, estimate the field strength at point D. Why is this an estimate?

35 Write an equation for the potential at point A due to Q_1 and Q_2. If the charge Q_1 is 1nC, find the value Q_2.

36 If an electron is moved between the following points, calculate the work done in eV (remember an electron is negative).
(a) E to A **(b)** C to F **(c)** A to C

6.7 Magnetic force and field

Assessment statements
6.3.1 State that moving charges give rise to magnetic fields.
6.3.2 Draw magnetic field patterns due to currents.
6.3.3 Determine the direction of the force on a current-carrying conductor in a magnetic field.
6.3.4 Determine the direction of the force on a charge moving in a magnetic field.
6.3.5 Define the magnitude and direction of a magnetic field.
6.3.6 Solve problems involving magnetic forces, fields and currents.



Not all magnets are man-made; certain rocks (for example, this piece of magnetite) are naturally magnetic.

What is a magnet?

We all know that magnets are the things that stick notes to fridge doors, but do we understand the forces that cause magnets to behave in this way?

There is evidence in ancient Greek and Chinese writing, that people knew about magnets more than 2600 years ago. We do not know who discovered the first magnet, but since antiquity it was known that if a small piece of magnetite was suspended on a string and held close to another larger piece, the small one experienced a turning force, causing it to rotate. It was also found that if the small rock was held on its own, it would always turn to point toward the North Pole. The end of the rock that pointed north was named the *north-seeking pole*, the other end was named the *south-seeking pole*.

Magnetic poles

The north-seeking pole (red) always points to the north.

Every magnet has two poles (north and south). A magnet is therefore called a *dipole*. It is not possible to have a single magnetic pole or monopole. This is not the same as electricity where you can have a dipole or monopoles. If you cut a magnet in half, each half will have both poles.

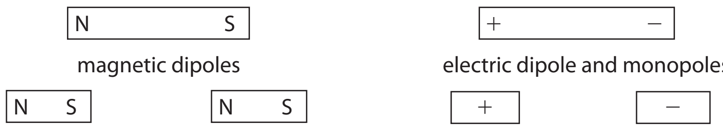


Figure 6.41 Magnets are dipoles.

Unlike poles attract

If we take two magnets and hold them next to each other, we find that the magnets will turn so that the S and N poles come together.

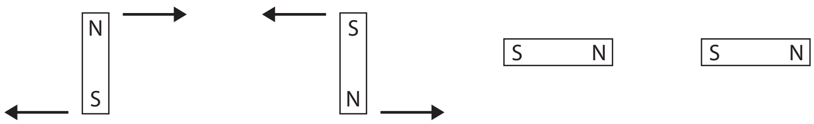


Figure 6.42 Magnets experience a turning force causing unlike poles to come together.

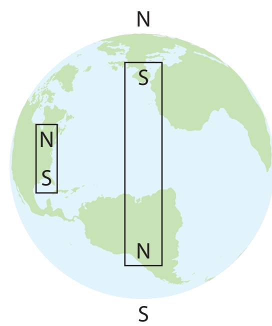
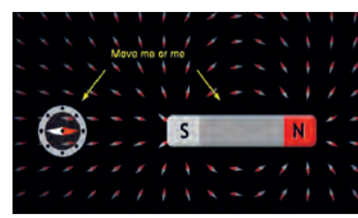


Figure 6.43 The north-seeking pole of a compass points north.

We can therefore conclude that the reason that a small magnet points toward the North Pole of the Earth is because there is a south magnetic pole there. This can be a bit confusing, but remember that the proper name for the pole of the magnet is north-seeking pole.

Magnetic field

Magnetism is similar to gravitational force and electric force in that the effect is felt even though the magnets do not touch each other; we can therefore use the concept of field to model magnetism. However, magnetism isn't quite the same; we described both gravitational and electric fields in terms of the force experienced by a small mass or charge. A small magnet placed in a field does not accelerate – it rotates, and therefore magnetic field is defined as a region of space where a small test magnet experiences a turning force.



W To plot magnetic fields on the PhET 'Faraday's electromagnetic lab', visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 6.5.

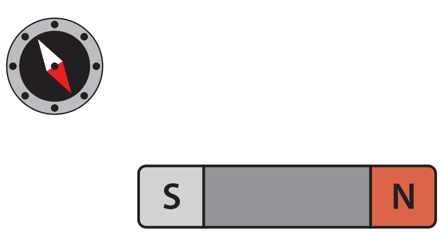


Figure 6.44 The small magnet is caused to turn, so must be in a magnetic field.

Since a small magnet rotates if held above the Earth, we can therefore conclude that the Earth has a magnetic field.

Magnetic field lines

In practice, a small compass can be used as our test magnet. Magnetic field lines are drawn to show the direction that the N pole of a small compass would point if placed in the field.

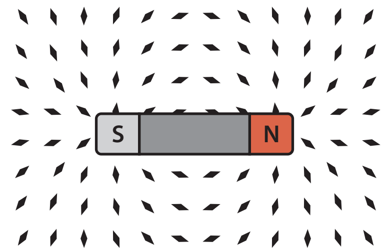


Figure 6.45 If the whole field were covered in small magnets, then they would show the direction of the field lines.

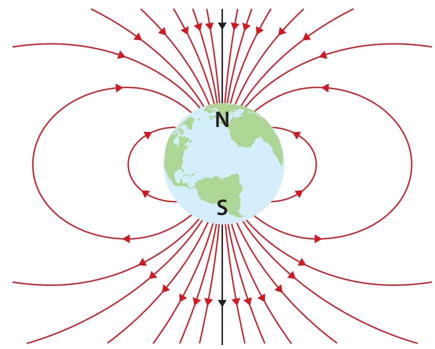


Figure 6.46 The Earth's magnetic field.

***B* field**
Since the letter *B* is used to denote flux density, the magnetic field is often called a *B field*.

Magnetic flux density (*B*)

From what we know about fields, the strength of a field is related to the density of field lines. This tells us that the magnetic field is strongest close to the poles. The magnetic flux density is the quantity that is used to measure how strong the field is – however it is not quite the same as field strength as used in gravitational and electric fields.

The unit of magnetic flux density is the tesla (T) and it is a vector quantity.

Field caused by currents

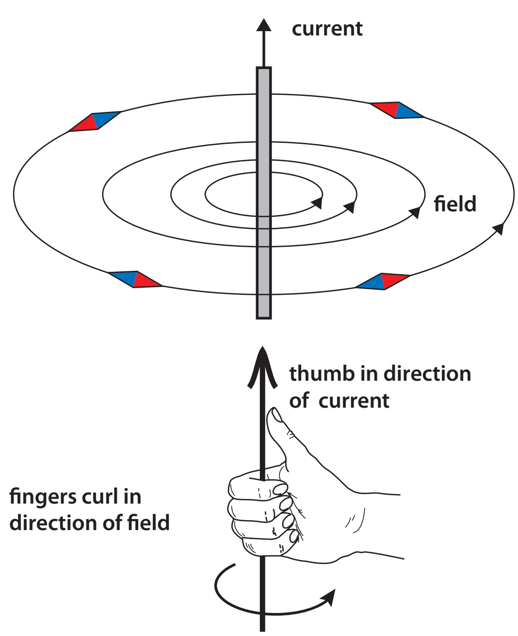


Figure 6.47 The field due to a long straight wire carrying a current is in the form of concentric circles so the field is strongest close to the wire.

If a small compass is placed close to a straight wire carrying an electric current, then it experiences a turning force that makes it always point around the wire. The region around the wire is therefore a magnetic field. This leads us to believe that magnetic fields are caused by moving charges.

Field inside a coil

When a current goes around a circular loop, the magnetic field forms circles.

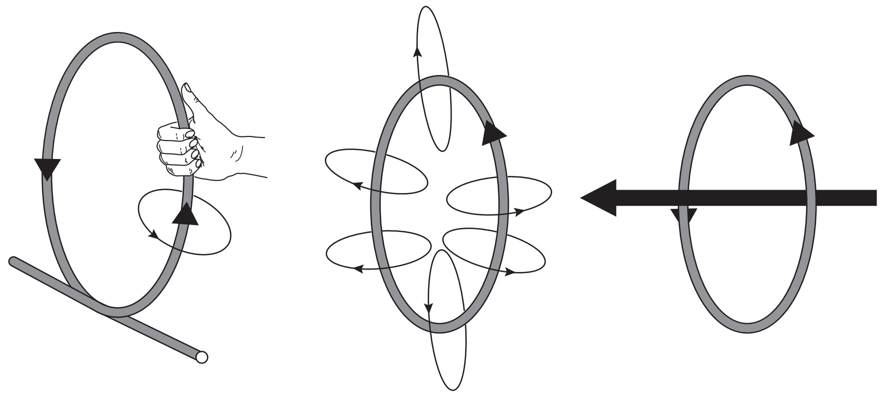


Figure 6.48 The direction of the field can be found by applying the right-hand grip rule to the wire. The circles formed by each bit of the loop add together in the middle to give a stronger field.

The field inside a solenoid

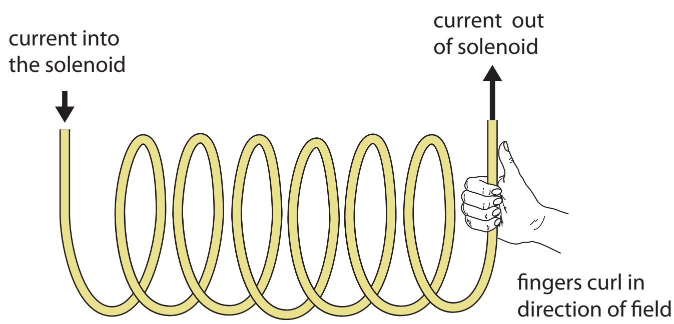


Figure 6.49 The direction of the field in a solenoid can be found using the grip rule on one coil.

The resulting field pattern is like that of a bar magnet but the lines continue through the centre.

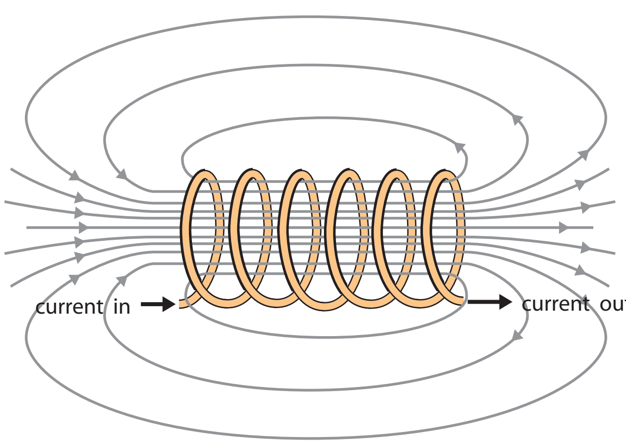
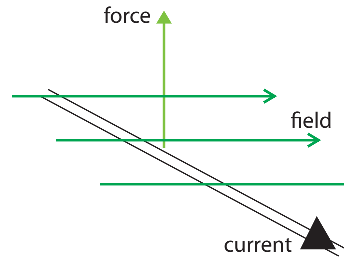


Figure 6.50 The field inside a solenoid.

Force on a current-carrying conductor

We have seen that when a small magnet is placed in a magnetic field, each end experiences a force that causes it to turn. If a straight wire is placed in a magnetic field, it also experiences a force. However, in the case of a wire, the direction of the force does not cause rotation – the force is in fact perpendicular to the direction of both current and field.

Figure 6.51 Force, field and current are at right angles to each other.



Definition of the ampere
The ampere is defined in terms of the force between two parallel current carrying conductors. A current of 1 A causes a force of 2×10^{-7} N per meter between two long parallel wires placed 1 m apart in a vacuum.

The size of this force is dependent on:

- how strong the field is – flux density B
- how much current is flowing through the wire – I
- the length of the wire – l

If B is measured in tesla, I in amps and l in metres,

$$F = BIl$$

Figure 6.52 Using Fleming's left hand rule to find the direction of the force.

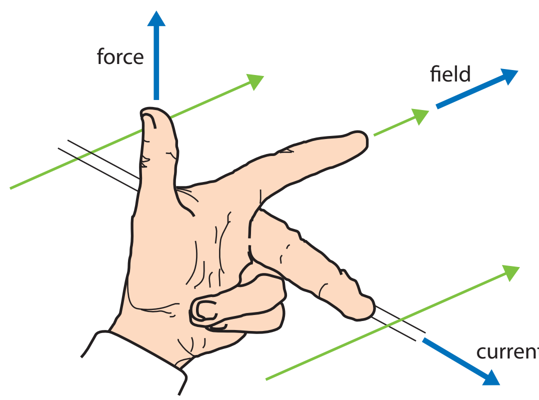
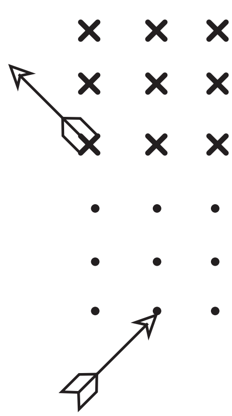


Figure 6.53 Field into the page can be represented by crosses, and field out by dots. Think what it would be like looking at an arrow from the ends.



Worked example

What is the force experienced by a 30 cm long straight wire carrying a 2 A current, placed in a perpendicular magnetic field of flux density 6 μT?

Solution

$B = 6\ \mu\text{T}$
$I = 2\ \text{A}$
$l = 0.3\ \text{m}$
$F = 6 \times 10^{-6} \times 2 \times 0.3\ \mu\text{N}$
$= 3.6\ \mu\text{N}$

● **Examiner's hint:** Use the formula $F = B \times I \times l$

Exercises

37 A straight wire of length 0.5 m carries a current of 2 A in a north–south direction. If the wire is placed in a magnetic field of 20 μT directed vertically downwards
(a) what is the size of the force on the wire?
(b) what is the direction of the force on the wire?

38 A vertical wire of length 1m carries a current of 0.5 A upwards. If the wire is placed in a magnetic field of strength 10 μT directed towards the N geographic pole
(a) what is the size of the force on the wire?
(b) what is the direction of the force on the wire?

Charges in magnetic fields

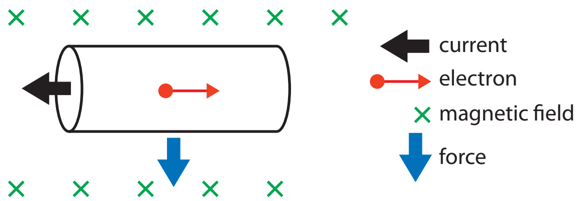


Figure 6.54 The force experienced by each electron is in the downward direction. Remember the electrons flow in the opposite direction to the conventional current.

From the microscopic model of electrical current, we believe that the current is made up of charged particles (electrons) moving through the metal. Each electron experiences a force as it travels through the magnetic field; the sum of all these forces gives the total force on the wire. If a free charge moves through a magnetic field, then it will also experience a force. The direction of the force is always perpendicular to the direction of motion, and this results in a circular path.

The force on each charge q moving with velocity v perpendicular to a field B is given by the formula $F = Bqv$.

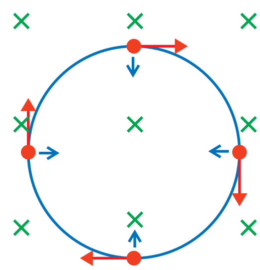


Figure 6.55 Wherever you apply Fleming's left hand rule, the force is always towards the centre.

Hint: Remember – the electron is negative, so current is in opposite direction to electron flow.

6.8 Electromagnetic induction

Assessment statements

12.1.1 Describe the inducing of an emf by relative motion between a conductor and a magnetic field.
12.1.2 Derive the formula for the emf induced in a straight conductor moving in a magnetic field.
12.1.3 Define magnetic flux and magnetic flux linkage.
12.1.4 Describe the production of an induced emf by a time-changing magnetic flux.
12.1.5 State Faraday's law and Lenz's law.
12.1.6 Solve electromagnetic induction problems.

Conductor moving in a magnetic field

We have considered what happens to free charges moving in a magnetic field, but what happens if these charges are contained in a conductor? Figure 6.56 shows a conductor of length L moving with velocity v through a perpendicular field of flux density B. We know from our microscopic model of conduction that conductors contain free electrons. As the free electron shown moves downwards through the field it will experience a force. Using Fleming's left hand rule, we can deduce that the direction of the force is to the left. (Remember, the electron is negative so if it is moving downwards the current is upwards.) This force will cause the free electrons to move to the left as shown in Figure 6.57. We can see that the electrons moving left have caused the lattice atoms on the right to become

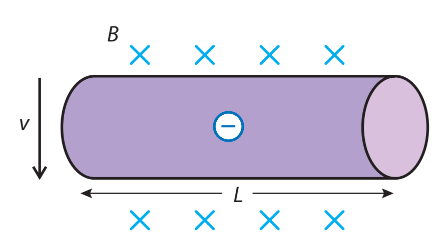
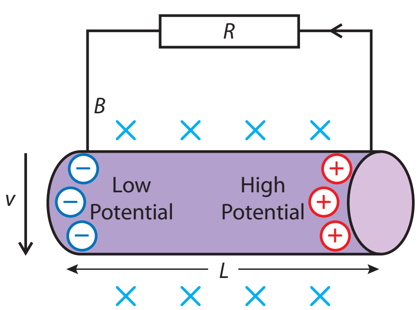


Figure 6.56 A conductor moving through a perpendicular field.

positive, and there is now a potential difference between the ends of the conductor. The electrons will now stop moving because the *B* force pushing them left will be balanced by an *E* force pulling them right.

Induced emf

Figure 6.57 Current flows from high potential to low potential.



If we connect this moving conductor to a resistor, a current would flow, as the moving wire is behaving like a battery. If current flows through a resistor then energy will be released as heat. This energy must have come from the moving wire. When a battery is connected to a resistor, the energy released comes from the chemical energy in the battery. We defined the amount of chemical energy converted to electrical energy per unit charge as the emf. We use the same quantity here but call it the induced emf.

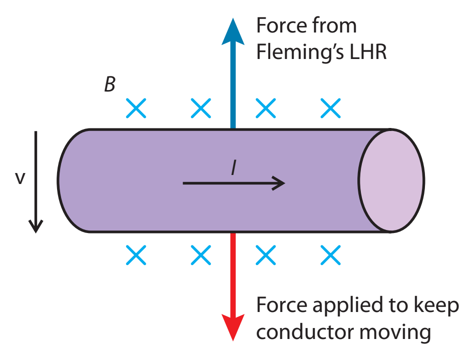
The emf is the amount of mechanical energy converted into electrical energy per unit charge.

The unit of induced emf is the volt.

Conservation of energy

When current flows through the resistor, current will also flow from left to right through the moving conductor. We now have a current carrying conductor in a magnetic field, allowing more electrons to move to the right. There is now a current flowing through the conductor from right to left. We now have a current-carrying conductor in a magnetic field, which, according to Fleming's left hand rule, will experience a force upwards (see Figure 6.58). This means that to keep it moving at constant velocity we must exert a force downwards, which means that we are now doing work. This work increases the electrical PE of the charges; when the charges then flow through the resistor this electrical PE is converted to heat, and energy is conserved.

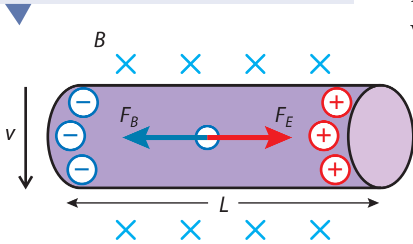
Figure 6.58 A constant force must be applied to the conductor to keep it moving.



Calculating induced emf

The maximum p.d. achieved across the conductor is when the magnetic force pushing the electrons left equals the electric force pushing them right. When the forces are balanced, no more electrons will move. Figure 6.59 shows an electron with balanced forces.

Figure 6.59 The electric and magnetic forces acting on the electron are balanced.



The second law of thermodynamics
You can see that when the electrons are pushed to the left of the conductor they are becoming more ordered. So that this does not break the second law, there must be some disorder taking place.

If F_B is the magnetic force and F_E is the electric force we can say that

$$F_B = F_E$$

Now we know that if the velocity of the electron is v and the field strength is B then $F_B = Bev$

The electric force is due to the electric field E which we can find from the equation $E = -\frac{dV}{dx}$ that we established in the section on electric potential. In this case, the field is uniform so the potential gradient $= \frac{V}{L}$

So, $$F_E = Ee = \frac{Ve}{L}$$

Equating the forces gives $$\frac{Ve}{L} = Bev$$

so $$V = BLv$$

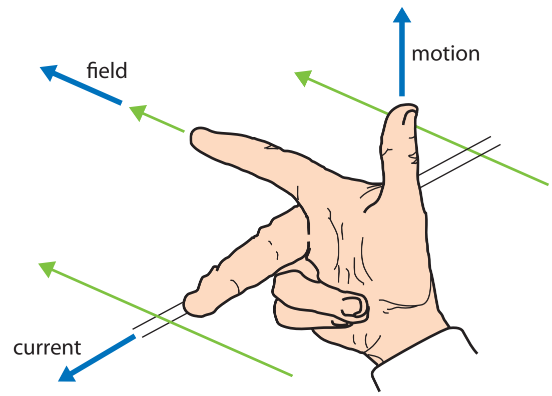
This is the p.d. across the conductor, which is defined as the work done per unit charge, taking a small positive test charge from one side to the other. As current starts to flow in an external circuit, the work done by the pulling force enables charges to move from one end to the other, so the emf (mechanical energy converted to electrical per unit charge) is the same as this p.d.

$$\text{induced emf} = BLv$$

● **Hint:**

Fleming's right hand rule

The fingers represent the same things as in the left hand rule but it is used to find the direction of induced current if you know the motion of the wire and the field. Try using it on the example in Figure 6.58.



Non-perpendicular field

If the field is not perpendicular to the direction of motion then you take the component of the flux density that is perpendicular. In the example in Figure 6.60, this is $B \sin\theta$.

So emf $= B\sin\theta \times Lv$

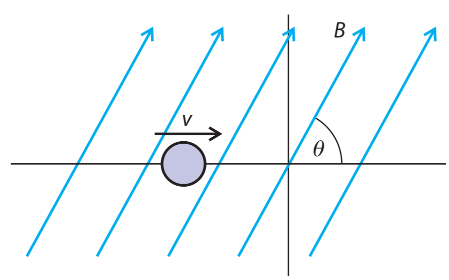


Figure 6.60 Looking along the wire, the component of the flux density that is perpendicular to the direction of motion is $B \sin\theta$.

Exercises

39 A 20 cm long straight wire is travelling at a constant 20 m s^{-1} through a perpendicular B field of flux density 50 μT.

(a) Calculate the emf induced.

(b) If this wire were connected to a resistance of 2 Ω how much current would flow?

(c) How much energy would be converted to heat in the resistor in 1s? (Power $= I^2R$)

(d) How much work would be done by the pulling force in 1s?

(e) How far would the wire move in 1s?

(f) What force would be applied to the wire?

Faraday's law

From the moving conductor example we see that the induced emf is dependent on the flux density, speed of movement and length of conductor. These three factors all change the rate at which the conductor cuts through the field lines, so a more convenient way of expressing this is:

The induced emf is equal to the rate of change of flux.

This is Faraday's law and it applies to all examples of induced emf not just wires moving through fields.

This can be written as $E = \frac{d\Phi}{dt}$

Flux and flux density

We can think of flux density as being proportional to the number of field lines per unit area, so flux is proportional to the number of field lines in a given area. If we take the example in Figure 6.61, the flux density of the field shown is B and the flux φ enclosed by the shaded area is BA.

The unit of flux is tesla metre² (Tm^2).

The wire in the same diagram will move a distance v in 1s (velocity is displacement per unit time) so the area swept out per unit time $= Lv$. The flux cut per unit time will therefore be BLv; this is equal to the emf.

If the field is not perpendicular to the area then you use the component of the field that is perpendicular. In the example in Figure 6.62 this would give

$$\varphi = B\cos\theta \times A$$

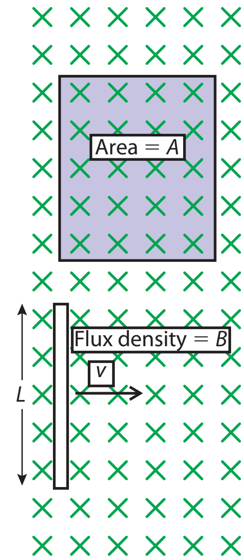


Figure 6.61 A wire is moving through a uniform B field

● **Hint:** If you think of the field lines as grass stems (they've been drawn green to help your imagination) and the conductor as a blade, then the induced emf is proportional to the rate of grass cut. This can be increased by moving the blade more quickly, having a longer blade or moving to somewhere where there is more grass.

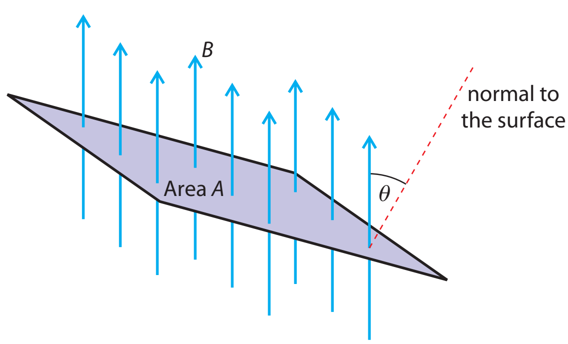


Figure 6.62 Area A not perpendicular to the field.

Lenz's law

We noticed that when a current is induced in a moving conductor, the direction of induced current causes the conductor to experience a force that opposes its motion. To keep the conductor moving will therefore require a force to be exerted in the opposite direction. This is a direct consequence of the law of conservation of energy. If it were not true, you wouldn't have to do work to move the conductor, so the energy given to the circuit would come from nowhere. Lenz's law states this fact in a way that is applicable to all examples:

The direction of the induced current is such that it will oppose the change producing it.

$$E = -\frac{dN\Phi}{dt}$$

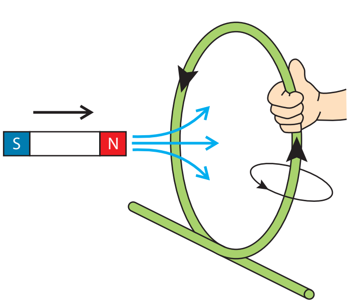


Figure 6.63 To oppose the magnet coming into the coil, the coil's magnetic field must push it out. The direction of the current is found using the grip rule.

Examples

1 **Coil and magnet**

A magnet is moved towards a coil as in Figure 6.63.

Applying Faraday's law:

As the magnet approaches the coil, the B field inside the coil increases, and the changing flux enclosed by the coil induces an emf in the coil that causes a current to flow. The size of the emf will be equal to the rate of change of flux enclosed by the coil.

Applying Lenz's law:
The direction of induced current will be such that it opposes the change producing it, which in this case is the magnet moving towards the coil. So to oppose this, the current in the coil must induce a magnetic field that pushes the magnet away; this direction is shown in the diagram.

2 Coil in a changing field
In Figure 6.64, the magnetic flux enclosed by coil B is changed by switching the current in coil A on and off.

Applying Faraday's law:
When the current in A flows, a magnetic field is created that causes the magnetic flux enclosed by B to increase. This increasing flux induces a current in coil B.

Applying Lenz's law:
The direction of the current in B must oppose the change producing it, which in this case is the increasing field from A. So to oppose this, the field induced in B must be in the opposite direction to the field from A, as in the diagram. This is the principle behind the operation of a transformer.

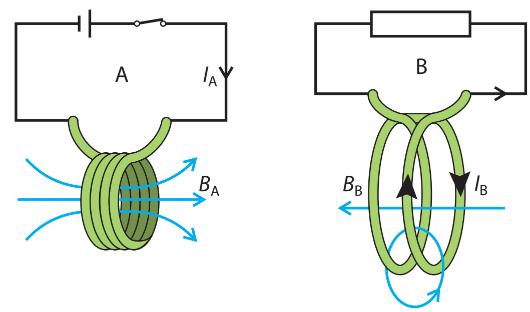


Figure 6.64 Coil A induces current in coil B. Use the grip rule to work out the direction of the fields.

Exercises

40 A coil with 50 turns and area 2 cm^2 encloses a field that is of flux density 100 μT (the field is perpendicular to the plane of the coil).
(a) What is the total flux enclosed?
(b) If the flux density changes to 50 μT in 2 s, what is the rate of change of flux?
(c) What is the induced emf?

41 A rectangular coil with sides 3 cm and 2 cm and 50 turns lies flat on a table in a region of magnetic field. The magnetic field is vertical and has flux density 500 μT.
(a) What is the total flux enclosed by the coil?
(b) If one side of the coil is lifted so that the plane of the coil makes an angle of 30° to the table, what will the new flux enclosed be?
(c) If the coil is lifted in 3 s, estimate the emf induced in the coil.

Hint: If a coil has N turns each turn of the wire encloses the flux. If the flux enclosed by the area of the coil is φ then the total flux enclosed is $N\varphi$.

Applications of electromagnetic induction

Induction braking

Traditional car brakes use friction pads that press against a disc attached to the road wheel. Induction braking systems replace the friction pads with electromagnets. When switched on, a current is induced in the rotating discs. According to Lenz's law the induced current will oppose the change producing it, resulting in a force that slows down the car.

In an induction braking system there are no parts to wear out, which means you would not have to replace brake pads every couple of years. This is bad news for the manufacturers of brake pads. Sometimes advancements in technology are not developed because they do not benefit manufacturers.

Induction cooking

An induction hotplate uses a changing magnetic field to induce an emf in a metal saucepan; the emf causes a current to flow in the saucepan, which produces heat (I^2R). The benefit of this system is that the saucepan (and therefore the food) gets hot but not the oven; you can touch the 'hotplate' without getting burned.

6.9 Alternating current

Assessment statements

12.2.1 Describe the emf induced in a coil rotating within a uniform magnetic field.
12.2.2 Explain the operation of a basic alternating current (AC) generator.
12.2.3 Describe the effect on the induced emf of changing the generator frequency.
12.2.4 Discuss what is meant by the root mean squared (rms) value of an alternating current or voltage.
12.2.5 State the relation between peak and rms values for sinusoidal currents and voltages.
12.2.6 Solve problems using peak and rms values.
12.2.7 Solve AC circuit problems for ohmic resistors.
12.2.8 Describe the operation of an ideal transformer.
12.2.9 Solve problems on the operation of ideal transformers.

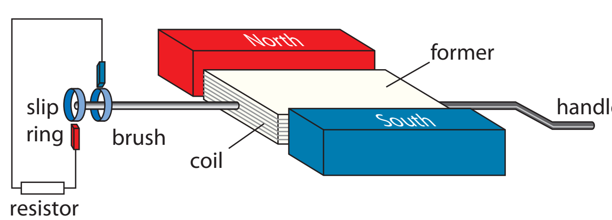
To view a simulation of an electric generator, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 6.6.

The most common application of electromagnetic induction is in the production of electrical energy. There are several devices that can be used to do this, such as the dynamo, in which a coil rotates in a magnetic field, or an alternator, in which a magnet rotates in a coil. Here we will consider the simple case of a coil rotating in a uniform magnetic field.

Coil rotating in a uniform magnetic field

Consider the coil shown in Figure 6.65. This coil is being made to rotate in a uniform magnetic field by someone turning the handle. The coil is connected to a resistor, but to prevent the wires connected to the coil twisting, they are connected via two slip rings. Resting on each slip ring is a carbon brush, which makes contact with the ring whilst allowing it to slip past.

Figure 6.65 A simple AC generator.



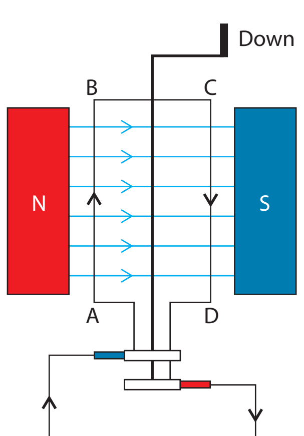


Figure 6.66 Looking at the generator from above.

To make the operation easier to understand, a simpler 2D version with only one loop of wire in the coil is shown in Figure 6.66. As the handle is turned, the wire on the left hand side (AB) moves up through the field. As it cuts the field a current will be induced. Using Fleming's right hand rule, we can deduce that the direction of the current is from A to B as shown. The direction of motion of the right hand side (CD) is opposite so the current is opposite. The result is a clockwise current through the resistor.

After turning half a revolution the coil is in the position shown in Figure 6.67. Side CD is now moving up through the field. Look carefully at how the slip ring has moved and you will see why, although the current is still clockwise in the coil, it is anticlockwise in the resistor circuit.

The size of the emf induced in a rotating coil

To find the size of the emf, we can use Faraday's law. This states that the induced emf will be equal to the rate of change of flux. The flux enclosed by the coil is related to the angle the coil makes with the field. Figure 6.68 shows a coil of N turns at time t. At this moment the normal to the plane of the coil makes an angle θ with the field.

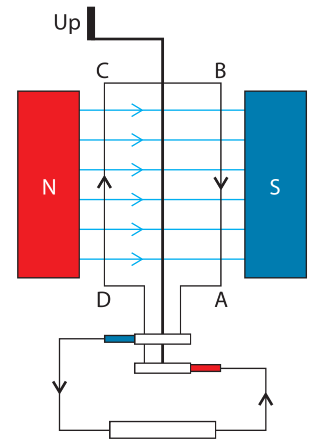


Figure 6.67 The coil after half a revolution.

Flux $\varphi = BA\cos\theta$

There are N turns so total flux $N\varphi = BAN\cos\theta$ (1)

If the angular velocity of the coil is ω then the angle $\theta = \omega t$

Substituting equation (1) gives

$$N\varphi = BAN\cos\omega t$$

This equation can be represented graphically as in Figure 6.69.

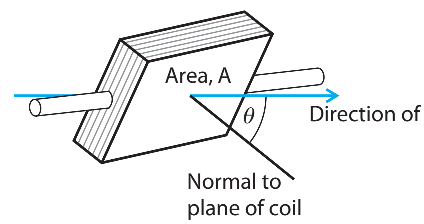


Figure 6.68 Coil at an angle to the field.

Note that the graph starts when the flux enclosed is a maximum, so $\cos\theta = 1$ meaning that $\theta = 0°$. Remember that θ is the angle between the field and the normal to the plane of the coil, so position A is when the coil in Figure 6.68 is vertical.

To find the magnitude of the emf, we need the rate of change of flux. We can calculate this from the gradient of this graph. Let us consider some specific points.

At A the gradient is zero.

At B the gradient is big and negative.

At C the gradient is zero.

At D the gradient is big and positive.

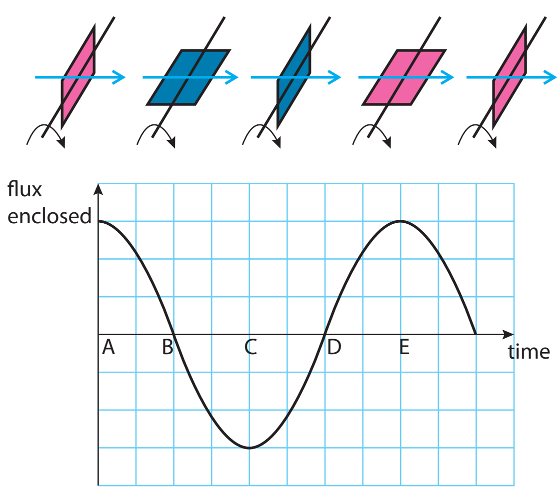


Figure 6.69 Graph of flux against time.

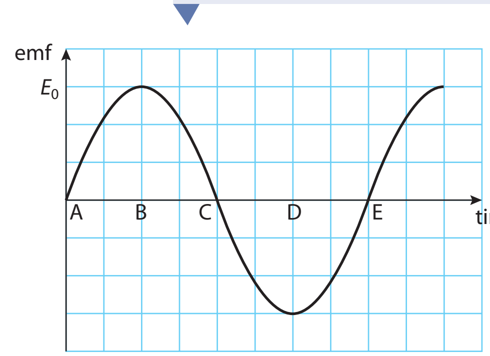
Lenz's law says that the induced current will oppose the change producing it; this means that when the rate change of flux is positive the induced emf is negative. If we plot the rate of change of flux against time we get a graph as in Figure 6.70. The equation of this line is

$$E = BAN\omega\sin\omega t$$

Where $BAN\omega$ is the peak value E_0

Note that this is a maximum when the flux enclosed is zero, that is when the coil is in the position shown in Figure 6.65.

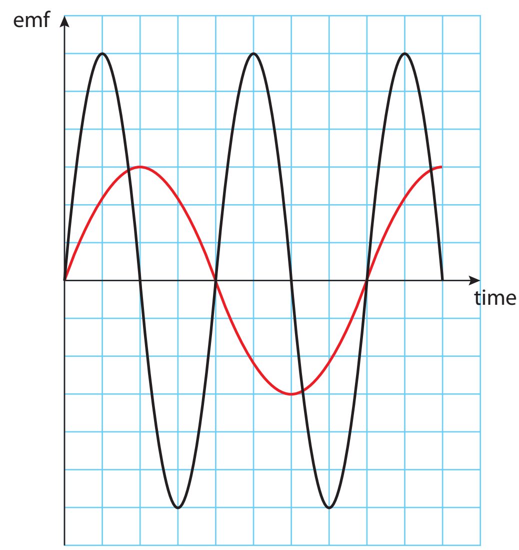
Figure 6.70 Graph of emf against time.



Effect of increasing angular speed

If the speed of rotation is increased, the graph of emf against time will change in two ways, as shown in Figure 6.71. Firstly, time between the peaks will be shorter, and secondly, the peaks will be higher. This is because if the coil moves faster, then the rate of change of flux will be higher and hence the emf will be greater.

Figure 6.71 The black graph is for a coil with twice the angular speed of the red one.



Faraday's law

If you have studied differentiation in maths you will understand that Faraday's law can be written

$$\epsilon = \frac{-dN\phi}{dt}$$

Then if $N\varphi = BAN \cos \omega t$

$$\frac{-dN\phi}{dt} = BAN\omega \sin \omega t$$

Alternating current

The current delivered by the rotating coil changes in direction and size over a period of time. This is called *alternating current* (AC). A battery, however, gives a constant current called *direct current* (DC). When doing power calculations with AC (for example, calculating the heat given out per second by an electric heater), the peak value would give a result that was too big since the peak value is only attained for a very short time. In these cases we should use the root mean square value; this is a sort of average value.

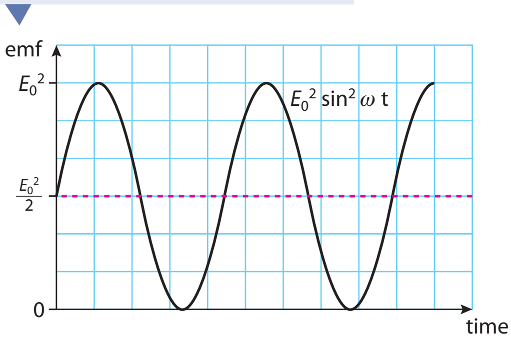
Root mean square

The root mean square or rms is the square root of the mean of the squares. Since the emf from the rotating coil varies sinusoidally then the rms emf and current will be the same as the root of the mean of the squares of the sine function.

To calculate the rms value, first we must square the function; this gives the curve shown in Figure 6.72. If we consider one complete cycle of this function we see that the mean value is $\frac{1}{2}E_0^2$. The rms value is then the square root of this.

$$E_{\text{rms}} = \sqrt{\frac{E_0^2}{2}} = \frac{E_0}{\sqrt{2}}$$

Figure 6.72



The current passing through a resistor will be proportional to the potential difference across it, so this will also be sinusoidal. If the peak current is I_0 the rms value will therefore be given by:

$$I_{\text{rms}} = \frac{I_0}{\sqrt{2}}$$

The rms values are what you need to know for power calculations, so it is these that are normally quoted. For example, the rms mains voltage in Europe = 220 V.

Power in AC circuits

If the rms current flowing through a resistor is I_{rms} and the rms potential difference across it is V_{rms} then the power dissipated in it will be $I_{rms} \times V_{rms}$.

Worked example

What will be the rms current flowing through a 100 W light bulb connected to a 220 V AC supply?

Solution

Power $= I_{rms}V_{rms}$

So $I_{rms} = \frac{\text{Power}}{V_{rms}} = \frac{100}{220} = 0.45\text{ A}$

> The electrical energy produced by a generator comes from the work done by the person turning the coil. The more current you take from the coil the harder it is to turn. This follows from Lenz's law; the current in the coil opposes the change producing it. If you don't draw any current then it's very easy to turn the coil.

Exercises

42 The rms voltage in the USA is 110 V. Calculate the peak voltage.

43 An electric oven designed to operate at 220 V has a power rating of 4 kW. What current flows through it when it is switched on?

44 A coil similar to the one in Figure 6.65 has an area of 5 cm^2 and rotates 50 times a second in a field of flux density 50 mT.

(a) If the coil has 500 turns, calculate

(i) the angular velocity, ω

(ii) the maximum induced emf

(iii) the rms emf.

(b) If the speed is reduced to 25 revolutions per second what is the new E_{rms}?

45 Calculate the resistance of a 1000 W light bulb designed to operate at 220 V.

The transformer

A transformer consists of two coils wound on a co mmon soft iron core as shown in Figure 6.73. The primary coil is connected to an AC supply. This causes a changing magnetic field inside the coil. This field is made stronger by the presence of the soft iron core which itself becomes magnetic (temporarily). Since the secondary is wound around the same former, it will have a changing magnetic field within it, which induces an emf in its coils.

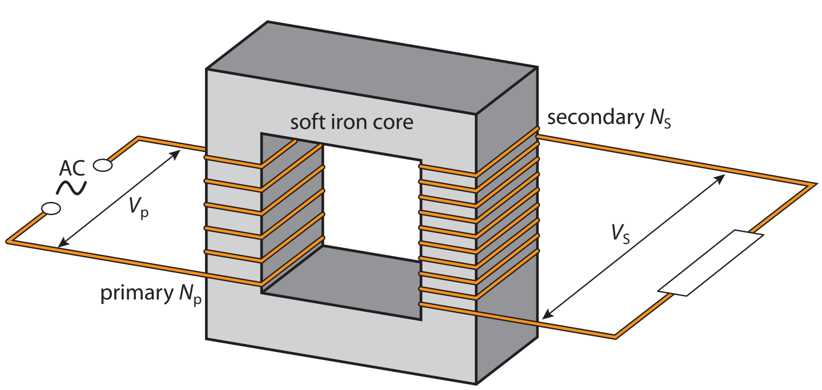


Figure 6.73 A simple transformer.

The emf induced in the secondary is directly related to the number of turns. If the supply is sinusoidal then the ratio of turns in the two coils equals the ratio of the p.d.s.

$$\frac{N_P}{N_S} = \frac{V_P}{V_S}$$

> **Power losses**
> Real transformers are not 100% efficient; power is lost in several ways:
> - Heat in the wire of the coils
> - Heat in the core due to induced currents in the soft iron.
> - Flux leakage.
> - The core retains some magnetism when the current changes.

So a transformer can make a p.d. higher or lower depending on the ratio of turns on the primary and secondary. However, since energy must be conserved, the power out cannot be bigger than the power in. Electrical power is given by $I \times V$ so if the p.d. goes up, the current must come down. An ideal transformer has an efficiency of 100%. This means that power in = power out.

$$V_P I_P = V_S I_S$$

where V_P, V_S, I_P and I_S are the rms values of p.d. and current in the primary and secondary.

Warning

The emf induced in the secondary depends on the rate of change of current in the primary. If you switch the current off then the change can be very big inducing a much bigger emf than you might have calculated.

Exercises

46 An ideal transformer steps down the 220 V mains to 4.5 V so it can be used to charge a mobile phone.

(a) If the primary has 500 turns how many turns does the secondary have?

(b) If the charger delivers 0.45 A to the phone, how much power does it deliver?

(c) How much current flows into the charger from the mains?

(d) The phone is unplugged from the charger but the charger is left plugged in. How much current flows into the charger now?

6.10 Transmission of electrical power

Assessment statements

12.3.1 Outline the reasons for power losses in transmission lines and real transformers.

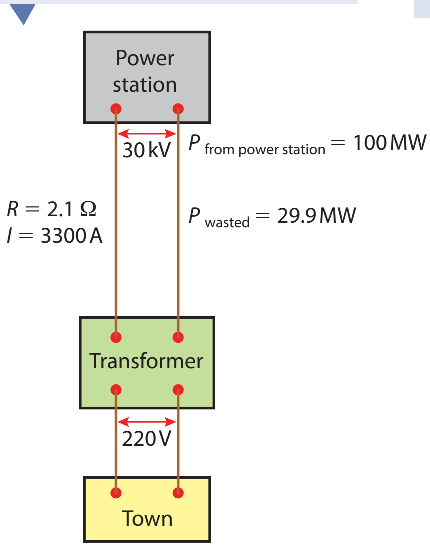
12.3.2 Explain the use of high-voltage step-up and step-down transformers in the transmission of electrical power.

12.3.3 Solve problems on the operation of real transformers and power transmission.

12.3.4 Suggest how extra-low-frequency electromagnetic fields, such as those created by electrical appliances and power lines, induce currents within a human body.

12.3.5 Discuss some of the possible risks involved in living and working near high-voltage power lines.

Figure 6.74



As outlined previously, electrical energy can be produced by moving a coil in a magnetic field. The energy required to rotate the coil can come from many sources, for example: coal, oil, falling water, sunlight, waves in the sea or nuclear fuel. The transformation of energy takes place in power stations that are often not sited close to the places where people live. For that reason, the electrical energy must be delivered via cables. These cables have resistance so some energy will be lost as the current flows through them. Let's take an example as illustrated in Figure 6.74.

The generators at a power station typically produce 100 MW of power at 30 kV. A small town 100 km from the power station requires electrical power which will be delivered via two aluminium cables (one there and one back) that have a radius of 2 cm. The houses in the town cannot use such high voltage electricity so the p.d. must be stepped down to 220 V, using a transformer that we will assume is 100% efficient.

The first thing to do is to calculate the resistance of the cables. We can do this using the formula

$$R = \frac{\rho l}{A}$$

where ρ is the resistivity of aluminium ($2.6 \times 10^{-8}\,\Omega$m), A is the cross-sectional area and l is the length.

● **Hint:** 30 kV is the p.d. between the two wires not the p.d. from the power station to the transformer.

So $$R = \frac{2.6 \times 10^{-8} \times 200 \times 10^3}{2\pi \times (2 \times 10^{-2})^2} = 2.1\,\Omega$$

The current that must be delivered if 100 MW is produced at a p.d. of 30 kV can be found using $P = IV$

Rearranging gives $$I = \frac{1 \times 10^8}{30 \times 10^3} = 3.3 \times 10^3\text{ A}$$

The power loss in the cables is therefore $I^2R = (3.3 \times 10^3)^2 \times 2.1 = 22.9$ MW

This is a lot of wasted power.

To reduce the power loss in the cables we must reduce the current. This can be done by stepping up the voltage before transmission, using a transformer as shown in the next example (Figure 6.75).

The p.d. between the wires is typically stepped up to 115 kV. Let us now repeat the calculation, assuming all transformers are 100% efficient.

The power is now delivered at 115 kV so the current is

$$I = \frac{P}{V} = \frac{100 \times 10^6}{115 \times 10^3} = 870\text{ A}$$

Power loss in the cables $= I^2R = 870^2 \times 2.1 = 1.6$ MW

This is still quite a lot of wasted power but much less than before.

To reduce this further, more cables could be added in parallel, thereby reducing the resistance.

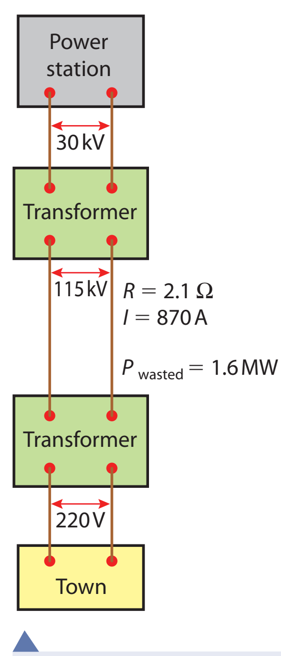


Figure 6.75

Health risks and power lines

Power lines carry large alternating currents at high potentials which produce electromagnetic fields radiating from them. The changing electric and magnetic field will induce currents in any conductors placed nearby. Your body is a conductor, so if you stand near a power line small currents will be induced in your body – is this harmful?

Unless you almost sit on the cable, these fields are very small. Typically the magnetic field you'd experience is less than the Earth's magnetic field. The frequency of the change is also very low so cannot affect the atoms of your body. It therefore seems very unlikely that any harm could be caused by these fields. In the 1970s, a study of the incidence of childhood leukaemia showed that there was a higher incidence of leukaemia in children living close to power lines. This caused a lot of concern at the time, but the study was seriously flawed. One of the main pieces of evidence against this theory is that even though the number of children living near power lines has increased significantly over the past 30 years, the incidence of leukaemia has gone down.

Power lines and leukaemia
This is a good example of how people can be influenced by the media. Even though there is no scientific evidence that power lines adversely affect one's health, people still believe what they heard on the TV 30 years ago. There are also a lot of companies marketing products that profit from this misconception.

Exercises

47 A power station that generates electricity at 50 kV produces 500 MW of power. This is delivered to a town through cables with a total resistance of 8Ω. Before transmission, the p.d. is stepped up to 100 kV, then stepped down to 220 V at the town.

(a) How much current will flow through the cables that take the electricity away from the power station to the town?

(b) How much power is lost in the cables?

(c) What percentage of total power delivered is lost?

(d) How much power will be delivered to the town? (Assume both transformers are 100% efficient.)

(e) How much power will be available for the town to use?

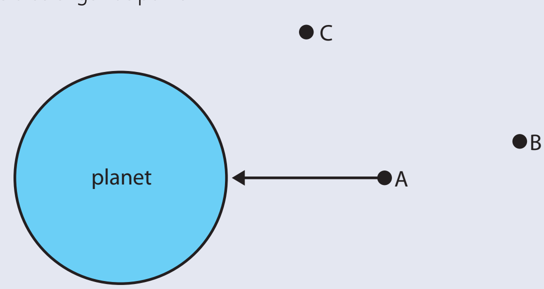
(f) How much total current will flow through the town?

Practice questions

1 This question is about gravitation and orbital motion.

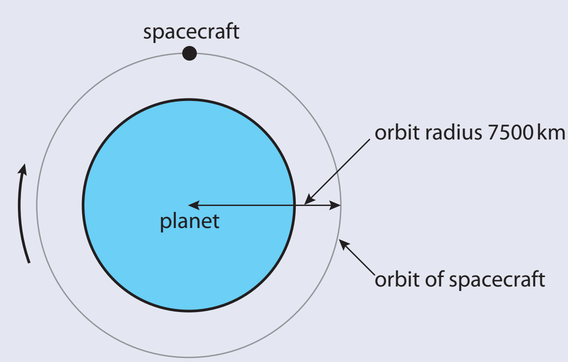
(a) Define gravitational field strength at a point in a gravitational field. (2)

The diagram below shows three points above a planet. The arrow represents the gravitational field strength at point A.



(b) Draw arrows to represent the gravitational field strength at point B and point C. (2)

A spacecraft is in a circular orbit around the planet as shown in the diagram below. The radius of the orbit is 7500 km.



(c) For the spacecraft in the position shown, draw and label arrows representing

(i) the velocity (label this arrow V). (1)

(ii) the acceleration (label this arrow A). (1)

The speed of the spacecraft is $6.5\ \text{km s}^{-1}$.

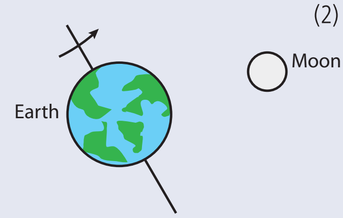
(d) Deduce the value of the magnitude of the gravitational field strength at a point in the spacecraft's orbit. (3)

2 This question is about gravitation and ocean tides.

(a) State Newton's law of universal gravitation. (2)

(b) Use the following information to deduce that the gravitational field strength at the surface of the Earth is approximately 10 N kg^{-1}.
Mass of the Earth $= 6.0 \times 10^{24}$ kg
Radius of the Earth $=$ 6400 km (2)

The Moon's gravitational field affects the gravitational field at the surface of the Earth. A high tide occurs at the point where the resultant gravitational field due to the Moon and to the Earth is a minimum.



(c) **(i)** On the diagram, label, using the letter P, the point on the Earth's surface that experiences the greatest gravitational attraction due to the Moon. Explain your answer. (2)

(ii) On the diagram label, using the letter H, the location of a high tide. Explain your answer. (2)

(iii) Suggest **two** reasons why high tides occur at different times of the day in different locations. (2)

3 This question is about gravitational fields.

(a) Define *gravitational field strength*. (2)

The gravitational field strength at the surface of Jupiter is 25 N kg^{-1} and the radius of Jupiter is 7.1×10^7 m.

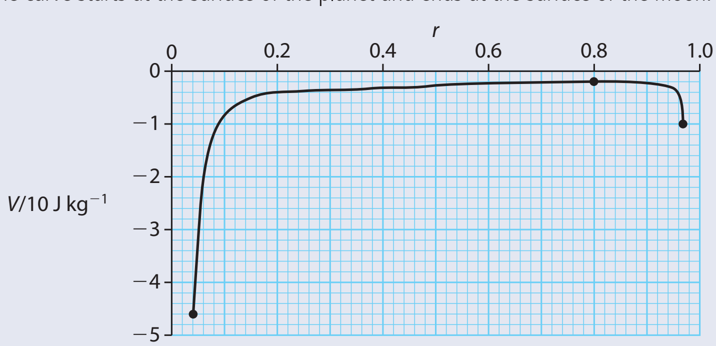
(b) **(i)** Derive an expression for the gravitational field strength at the surface of a planet in terms of its mass M, its radius R and the gravitational constant G. (2)

(ii) Use your expression in **(b) (i)** above to estimate the mass of Jupiter. (2)

4 This question is about gravitation.

(a) Define *gravitational potential* at a point. (2)

(b) The diagram below shows the variation of gravitational potential V of a planet and its moon with distance r from the centre of the planet. The unit of separation is arbitrary. The centre of the planet corresponds to $r = 0$ and the centre of the moon to $r = 1$. The curve starts at the surface of the planet and ends at the surface of the moon.



(i) At the position where $r = 0.8$, the gravitational field strength is zero. Determine the ratio

$$\frac{\text{mass of planet}}{\text{mass of moon}}$$ (3)

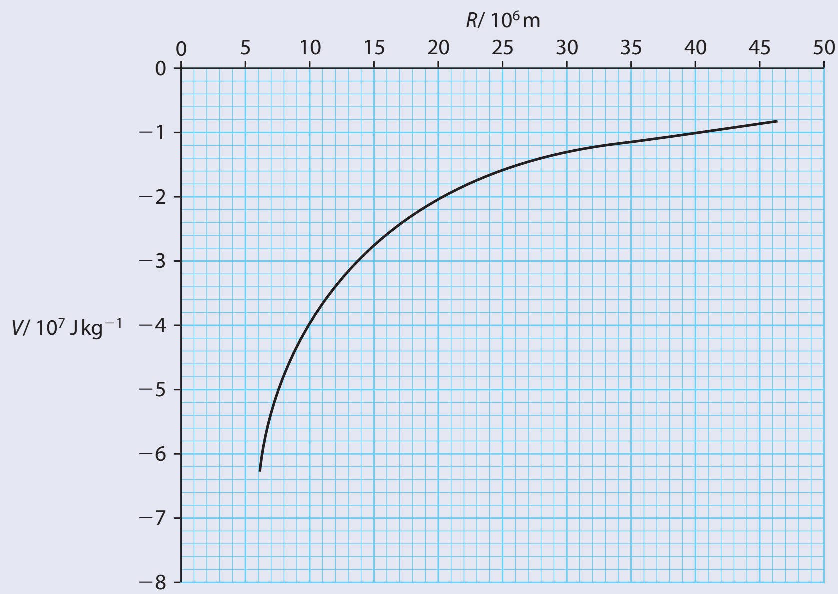
(ii) A satellite of mass 1500 kg is launched from the surface of the planet. Determine the **minimum** kinetic energy at launch the satellite must have so that it can reach the surface of the moon. (3)

(Total 8 marks)

5 This question is about gravitational potential energy.

The graph below shows the variation of gravitational potential V due to the Earth with distance R from the centre of the Earth. The radius of the Earth is 6.4×10^6 m.

The graph does not show the variation of potential V within the Earth.



(a) Use the graph to find the gravitational potential

(i) at the surface of the Earth. (1)

(ii) at a height of 3.6×10^7 m above the surface of the Earth. (2)

(b) Use the values you have found in part **(a)** to determine the minimum energy required to put a satellite of mass 1.0×10^4 kg into an orbit at a height of 3.6×10^7 m above the surface of the Earth. (3)

(c) Give **two** reasons why more energy is required to put this satellite into orbit than that calculated in **(b)** above.

(Total 8 marks)

6 This question is about the possibility of generating electrical power using a satellite orbiting the Earth.

(a) Define *gravitational field strength*. (2)

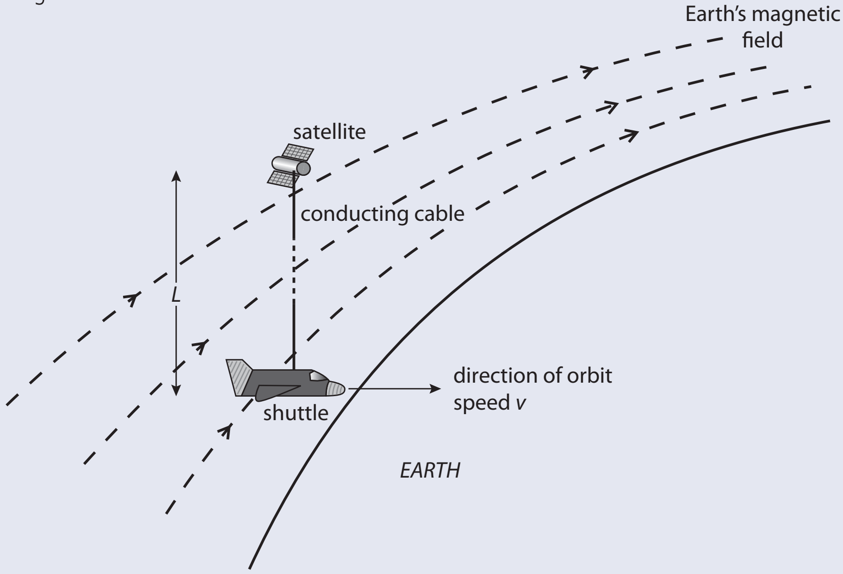
(b) Use the definition of gravitational field strength to deduce that

$$GM = g_0R^2$$

where M is the mass of the Earth, R its radius and g_0 is the gravitational field strength at the surface of the Earth. (You may assume that the Earth is a uniform sphere with its mass concentrated at its centre.) (2)

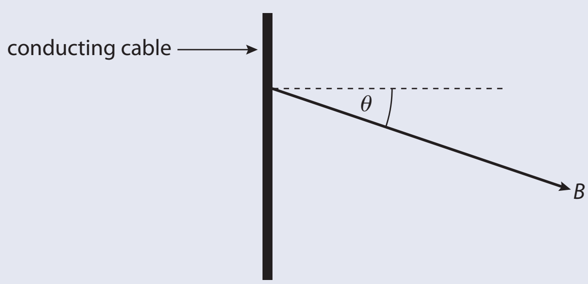
A space shuttle orbits the Earth and a small satellite is launched from the shuttle. The satellite carries a conducting cable connecting the satellite to the shuttle. When the satellite is a distance L from the shuttle, the cable is held straight by motors on the satellite.

Diagram 1



As the shuttle orbits the Earth with speed v, the conducting cable is moving at right angles to the Earth's magnetic field. The magnetic field vector $\boldsymbol{B}$ makes an angle θ to a line perpendicular to the conducting cable as shown in diagram 2. The velocity vector of the shuttle is directed out of the plane of the paper.

Diagram 2



(c) On diagram 2, draw an arrow to show the direction of the magnetic force on an electron in the conducting cable. Label the arrow F. (1)

(d) State an expression for the force F on the electron in terms of B, v, e and θ, where B is the magnitude of the magnetic field strength and e is the electron charge. (1)

(e) Hence deduce an expression for the emf E induced in the conducting wire. (3)

(f) The shuttle is in an orbit that is 300 km above the surface of the Earth. Using the expression

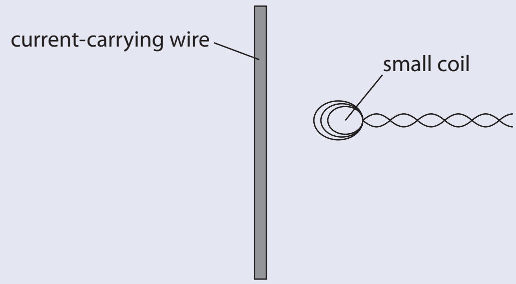
$$GM = g_0R^2$$

and given that $R = 6.4 \times 10^6$ m and $g_0 = 10\ \mathrm{N\,kg^{-1}}$, deduce that the orbital speed v of the satellite is $7.8 \times 10^3\ \mathrm{m\,s^{-1}}$. (3)

(g) The magnitude of the magnetic field strength is 6.3×10^{-6} T and the angle $\theta = 20°$. Estimate the length L of the cable required in order to generate an emf of 1 kV. (2)

(Total 14 marks)

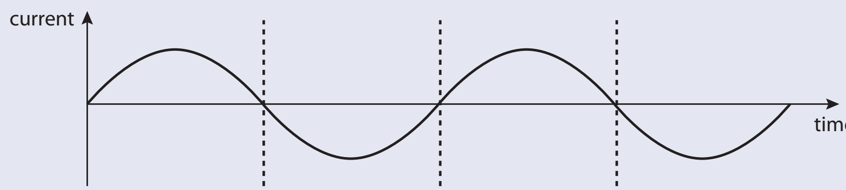
7 This question is about electromagnetic induction.
A small coil is placed with its plane parallel to a long straight current-carrying wire, as shown below.



(a) **(i)** State Faraday's law of electromagnetic induction. (2)

(ii) Use the law to explain why, when the current in the wire changes, an emf is induced in the coil. (1)

The diagram below shows the variation with time *t* of the current in the wire.



(b) **(i)** Draw, on the axes provided, a sketch-graph to show the variation with time *t* of the magnetic flux in the coil. (1)

(ii) Construct, on the axes provided, a sketch-graph to show the variation with time *t* of the emf induced in the coil. (2)

(iii) State and explain the effect on the maximum emf induced in the coil when the coil is further away from the wire. (2)

(c) Such a coil may be used to measure large alternating currents in a high-voltage cable. Identify **one** advantage and **one** disadvantage of this method. (2)

(Total 10 marks)

8 A resistor is connected in series with an alternating current supply of negligible internal resistance. The **peak value** of the supply voltage is V_0 and the **peak value** of the current in the resistor is I_0. The **average power** dissipation in the resistor is

A $\dfrac{V_0 I_0}{2}$ B $\dfrac{V_0 I_0}{\sqrt{2}}$

C $V_0 I_0$ D $2V_0 I_0$ (1)

9 The rms voltages across the primary and secondary coils in an ideal transformer are V_p and V_s respectively. The currents in the primary and secondary coils are I_p and I_s respectively.
Which **one** of the following statements is always true?

A $V_s = V_p$ B $I_s = I_p$

C $V_s I_s = V_p I_p$ D $\dfrac{V_s}{V_p} = \dfrac{I_s}{I_p}$ (1)

7 Atomic and nuclear physics

7.1 Atomic structure

Assessment statements

7.1.1 Describe a model of the atom that features a small nucleus surrounded by electrons.
7.1.2 Outline the evidence that supports a nuclear model of the atom.
7.1.3 Outline one limitation of the simple model of the nuclear atom.
7.1.4 Outline evidence for the existence of atomic energy levels.

13.1.8 Outline a laboratory procedure for producing and observing atomic spectra.
13.1.9 Explain how atomic spectra provide evidence for the quantization of energy in atoms.
13.1.10 Calculate wavelengths of spectral lines from energy level differences and vice versa.

The arrangement of charge in the atom

We already know that matter is made up of particles (atoms) and we used this model to explain the thermal properties of matter. We also used the idea that matter contains charges to explain electrical properties. Since matter contains charge and is made of atoms, it seems logical that atoms must contain charge. But how is this charge arranged?

What does it mean when we say we know these things? Do we know that this is true, or is it just the model that's true?

There are many possible ways that charges could be arranged in the atom, but since atoms are not themselves charged, they must contain equal amounts of positive and negative. Maybe half the atom is positive and half negative, or perhaps the atom is made of two smaller particles of opposite charge?

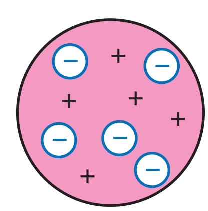


Figure 7.1 Thomson's model, positive pudding with negative plums.

The discovery of the electron by J.J. Thomson in 1897 added a clue that helps to solve the puzzle. The electron is a small negative particle that is responsible for carrying charge when current flows in a conductor. By measuring the charge-to-mass ratio of the electron, Thomson realised that the electrons were very small compared to the whole atom. He therefore proposed a possible arrangement for the charges as shown in Figure 7.1; this model was called the 'plum pudding' model. This model was accepted for some time until, under the direction of Ernest Rutherford, Geiger and Marsden performed an experiment that proved it could not be correct.

The electron is a fundamental particle with charge -1.6×10^{-19} C and mass 9.1×10^{-31} kg.

The Rutherford model

Rutherford's idea was to shoot alpha particles at a very thin sheet of gold to see what would happen. In 1909, when this was happening, very little was known

This is a good example of scientific method in practice.

To see a simulation of Rutherford scattering, visit www.heinemann.co.uk, enter the express code 4426P and click on weblink 7.1.

about alpha particles – only that they were fast and positive. In accordance with normal scientific practice, Rutherford would have applied the model of the day so as to predict the result of the experiment. The current model was that the atom was like a small plum pudding, so a sheet of gold foil would be like a wall of plum puddings, a few puddings thick. Shooting alpha particles at the wall would be like firing bullets at a wall of plum puddings. If we think what would happen to the bullets and puddings, it will help us to predict what will happen to the alpha particles.

If you shoot a bullet at a plum pudding, it will pass straight through and out the other side (you can try this if you like). What actually happened was, as expected, most alpha particles passed through without changing direction, but a significant, number were deflected and a few even came right back, as shown in Figure 7.2. This was so unexpected that Rutherford said 'It was quite the most incredible event that ever happened to me in my life. It was almost as incredible as if you had fired a 15-inch shell at a piece of tissue paper and it came back and hit you.' We know from our study of collisions that you can only get a ball to bounce off another one if the second ball is much heavier than the first. This means that there must be something heavy in the atom. The fact that most alphas pass through means that there must be a lot of space. If we put these two findings together, we conclude that the atom must have something heavy and small within it. This small, heavy thing isn't an electron since they are too light; it must therefore be the positive part of the atom. This would explain why the alphas come back, since they are also positive and would be repelled from it.

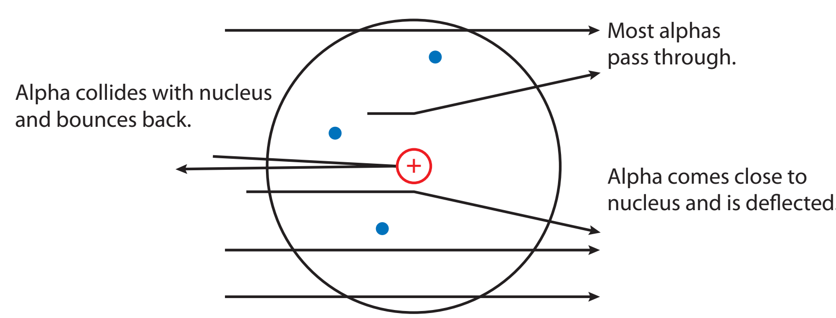


Figure 7.2 Alpha particles deflected by a gold atom.

Some chemistry
An atom consists of a nucleus surrounded by a number of electrons. Different elements have different-sized nuclei and different numbers of electrons. To make the atom neutral the nucleus has the same charge (but positive) as all the electrons put together. The number of electrons in the neutral atom is equal to the atomic number of the element and defines the way that it combines with other elements.

The Bohr model

In 1913 Niels Bohr suggested that the nucleus could be like a mini solar system with the electrons orbiting the nucleus. This fits in very nicely with what we know about circular motion, since the centripetal force would be provided by the electric attraction between the electron and nucleus. The problem with this model is that as the electrons accelerate around the nucleus, they would continually radiate electromagnetic radiation; they would therefore lose energy causing them to spiral into the nucleus. This model therefore cannot be correct. One thing that the Bohr model did come close to explaining was the line spectrum for hydrogen.

The connection between atoms and light

There is a very close connection between matter and light. For example, if we give thermal energy to a metal, it can give out light. Light is an electromagnetic wave so must come from a moving charge; electrons have charge and are able to move, so it would be reasonable to think that the production of light is something to do with electrons. But what is the mechanism inside the atom that enables this to happen? Before we can answer that question we need to look more closely at the nature of light, in particular light that comes from isolated atoms. We must look at isolated atoms because we need to be sure that the light is coming from single atoms and not the interaction between atoms. A single atom would not produce enough light for us to see, but low-pressure gases have enough atoms far enough apart not to interact.

Atomic spectra

To analyse the light coming from an atom we need to first give the atom energy; this can be done by adding thermal energy or electrical energy. The most convenient method is to apply a high potential to a low-pressure gas contained in a glass tube (a discharge tube). This causes the gas to give out light, and already you will notice (see Figure 7.3) that different gases give different colours. To see exactly which wavelengths make up these colours we can split up the light using a prism (or diffraction grating). To measure the wavelengths we need to know the angle of refraction; this can be measured using a spectrometer.

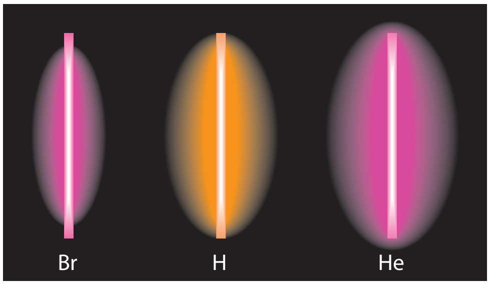


Figure 7.3 Discharge tubes containing bromine, hydrogen and helium.

The hydrogen spectrum

Hydrogen has only one electron – so it is the simplest atom and the one we will consider first. Figure 7.4 shows the spectrum obtained from a low-pressure discharge tube containing hydrogen. The first thing you notice is that, unlike a usual rainbow that is continuous, the hydrogen spectrum is made up of thin lines. Light is a form of energy, so whatever the electrons do they must lose energy when light is emitted. If the colour of light is associated with different energies, then, since only certain energies of light are emitted,the electron must only be able to release certain amounts of energy. This would be the case if the electron could only have certain amounts of energy in the first place. We say the energy is *quantized*.

Figure 7.4 The line spectrum for hydrogen.

To help understand this, we can consider an analogous situation of buying sand. You can either buy sand loose or in 50 kg bags, and we say the 50 kg bags are quantized, since the sand comes in certain discrete quantities. So if you buy loose sand, you can get any amount you want, but if you buy quantized sand, you have to have multiples of 50 kg. If we make a chart showing all the possible quantities of sand you could buy, then they would be as shown on Figure 7.5; one is continuous and the other has lines.

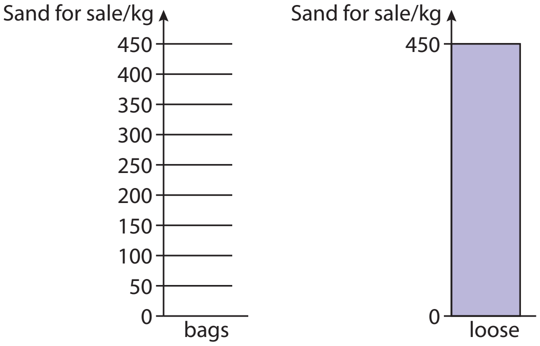


Figure 7.5 Ways of buying sand.

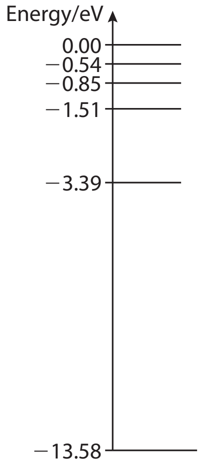


Figure 7.6 The electron energy levels of hydrogen.

If the electron in the hydrogen atom can only have discrete energies, then when it changes energy, that must also be in discrete amounts. We represent the possible energies on an energy level diagram (Figure 7.6), which looks rather like the sand diagram.

For this model to fit together, each of the lines in the spectrum must correspond to a different energy change. Light therefore must be quantized and this does not tie in at all with our classical view of light being a continuous wave that spreads out like ripples in a pond.

7.2 The quantum nature of light

Assessment statements

13.1.1 Describe the photoelectric effect.
13.1.2 Describe the concept of the photon, and use it to explain the photoelectric effect.
13.1.3 Describe and explain an experiment to test the Einstein model.
13.1.4 Solve problems involving the photoelectric effect.

Light definitely has wave-like properties; it reflects, refracts, diffracts and interferes. But sometimes light does things that we don't expect a wave to do, and one of these is the photoelectric effect.

The photoelectric effect

Consider ultraviolet light shining on a negatively charged zinc plate. Can the light cause the charge to leave the plate? To answer this question we can use the wave model of light, but we cannot see what is happening inside the metal, so to help us visualize this problem we will use an analogy.

The swimming pool analogy

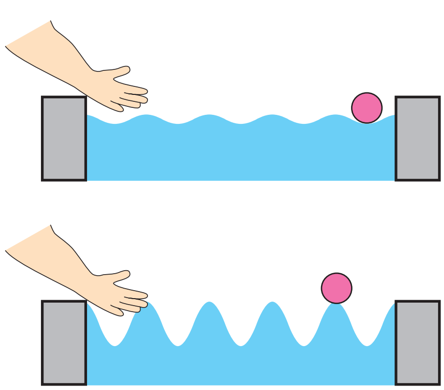


Figure 7.7

Imagine a ball floating near the edge of a swimming pool as in Figure 7.7. If you are at the other side of the swimming pool, could you get the ball out of the pool by sending water waves towards it? To get the ball out of the pool we need to lift it as high as the edge of the pool; in other words, we must give it enough PE to reach this height. We can do this by making the amplitude of the wave high enough to lift the ball. If the amplitude is not high enough, the ball will not be able to leave the pool unless we build a machine (as in Figure 7.8) that will collect the energy over a period of time. In this case, there will be a time delay before the ball gets out.

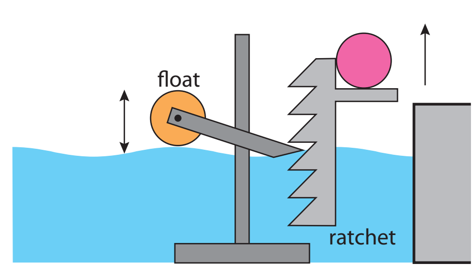


Figure 7.8

To relate this to the zinc plate, according to this model we expect:

- Electrons will be emitted only if the light source is very bright. (Brightness is related to amplitude of the wave.)
- If the source is dim we expect no electrons to be emitted. If electrons are emitted, we expect a time delay whilst the atoms collect energy from the wave.
- If we use lower frequency light, electrons will still be emitted if it is bright enough.

The zinc plate experiment

To find out if electrons are emitted or not, we can put the zinc plate on an electroscope and shine UV light on it as in Figure 7.9. If electrons are emitted charge will be lost and the electroscope leaf will fall. The results are not entirely as expected:

- The electroscope does go down indicating that electrons are emitted from the surface of the zinc plate.
- The electroscope leaf goes down even if the UV light is very dim. When very dim, the leaf takes longer to go down but there is no time delay before it starts to drop.
- If light of lower frequency is used, the leaf does not come down, showing that no electrons are emitted. This is the case even if very intense low frequency light is used.

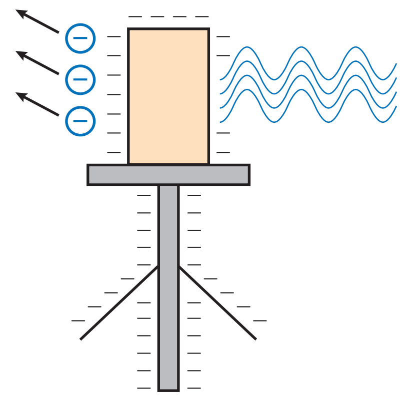


Figure 7.9 UV radiation is absorbed and electrons are emitted causing the gold leaf to fall.

These results can be explained if we consider light to be quantized.

Quantum model of light

In the quantum model of light, light is made up of packets called photons. Each photon is a section of wave with an energy E that is proportional to its frequency, f.

$$E = hf$$

where h is Planck's constant (6.63×10^{-34} J s).

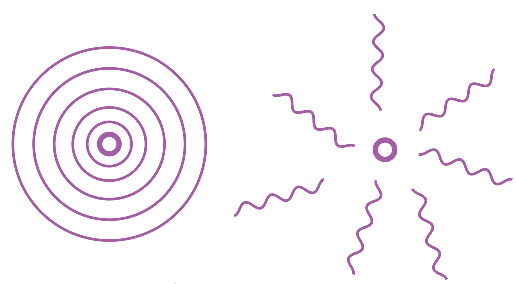


Figure 7.10 Light radiating, according to the wave model and the quantum model.

The intensity of light is related to the number of photons, not the amplitude, as in the classical wave model. Using this model we can explain the photoelectric effect:

- UV light has a high frequency, so when photons of UV light are absorbed by the zinc, they give enough energy to the zinc electrons to enable them to leave the surface.
- When the intensity of light is low there are not many photons but each photon has enough energy to enable electrons to leave the zinc.
- Low frequency light is made of photons with low energy; these do not have enough energy to liberate electrons from the zinc. Intense low frequency light is simply more low energy photons, which still don't have enough energy.

If a swimming pool were like this then if someone jumped into the pool the energy they give to the water would stay together in a packet until it met another swimmer. When this happened the other swimmer would be ejected from the pool. Could be fun!

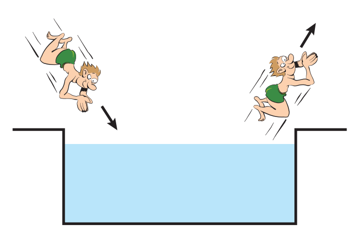


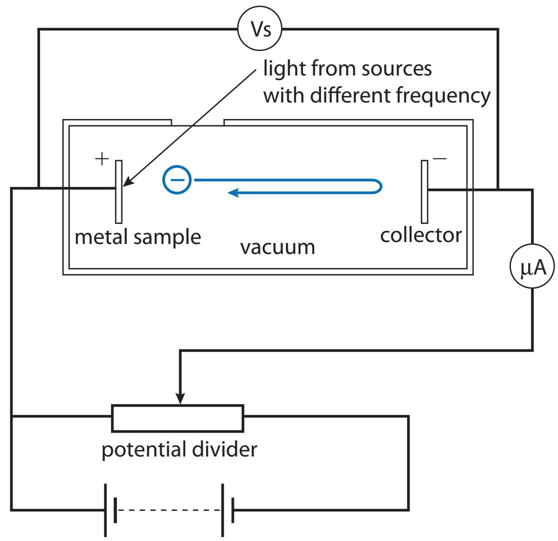
Figure 7.11 The quantum swimming pool.

To get a deeper understanding of the photoelectric effect we need more information about the energy of the photoelectrons.

Millikan's photoelectric experiment

Millikan devised an experiment to measure the KE of photoelectrons. He used an electric field to stop the electrons completing a circuit and used that 'stopping potential' to calculate the KE. A diagram of the apparatus is shown in Figure 7.12.

Figure 7.12 The stopping potential stops the electrons from reaching the collector.



For a simulation of the photoelectric effect, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on weblink 7.2.

Light from a source of known frequency passes into the apparatus through a small window. If the photons have enough energy, electrons will be emitted from the metal sample. Some of these electrons travel across the tube to the collector causing a current to flow in the circuit; this current is measured by the microammeter. The potential divider is now adjusted until none of the electrons reaches the collector (as in the diagram). We can now use the law of conservation of energy to find the KE of the fastest electrons.

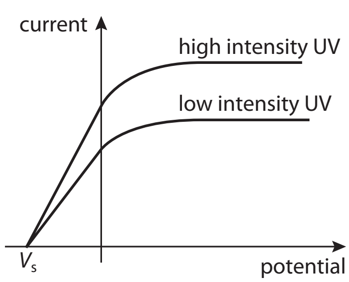
Loss of KE = gain in electrical PE

$$\tfrac{1}{2}mv^2 = V_s e$$

So maximum KE, $\mathrm{KE_{max}} = V_s e$

The light source is now changed to one with different frequency and the procedure is repeated.

The graphs in Figure 7.13 show two aspects of the results.



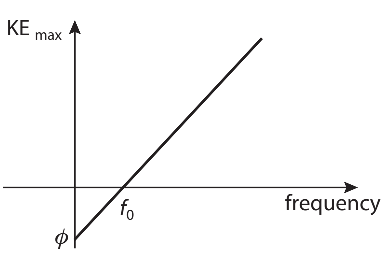


Figure 7.13 Graphs of current vs potential and maximum KE against frequency

The most important aspect of graph 1 is that for a given potential, increasing the intensity increases the current but doesn't change the stopping potential. This is because when the intensity is increased the light contains more photons (so liberates more electrons) but does not increase the energy (so V_s is the same).

Graph 2 shows that the maximum KE of the electrons is proportional to the frequency of the photons. Below a certain value, f_0, no photoelectrons are liberated; this is called the *threshold frequency*.

- **Examiner's hint:** Exam questions can include many different graphs related to the photoelectric effect. Make sure you look at the axis carefully.

Einstein's photoelectric equation

Einstein explained the photoelectric effect and derived an equation that relates the KE of the photoelectrons to the frequency of light.

Maximum photoelectron KE = energy of photon − energy needed to get photon out of metal

$$\mathrm{KE_{max}} = hf - \phi$$

ϕ is called the *work function*. If the photon has only enough energy to get the electron out then it will have zero KE, and this is what happens at the threshold frequency, f_0. At this frequency

$$KE_{max} = 0 = hf_0 - \phi$$

so

$$\phi = hf_0$$

We can now rewrite the equation as

$$KE_{max} = hf - hf_0$$

Equation of a straight line
The equation of a straight line is of the form $y = mx + c$ where m is the gradient and c is the y intercept. You should be able to see how the equation $KE_{max} = hf - \phi$ is the equation of the line in graph 2.

Exercises

1 A sample of sodium is illuminated by light of wavelength 422 nm in a photoelectric tube. The potential across the tube is increased to 0.6 V. At this potential no current flows across the tube. Calculate:

(a) the maximum KE of the photoelectrons
(b) the frequency of the incident photons
(c) the work function of sodium
(d) the lowest frequency of light that would cause photoelectric emission in sodium.

2 A sample of zinc is illuminated by UV light of wavelength 144 nm. If the work function of zinc is 4.3 eV, calculate

(a) the photon energy in eV
(b) the maximum KE of photoelectrons
(c) the stopping potential
(d) the threshold frequency.

3 If the zinc in Question 2 is illuminated by the light in Question 1, will any electrons be emitted?

4 The maximum KE of electrons emitted from a nickel sample is 1.4 eV. If the work function of nickel is 5.0 eV, what frequency of light must have been used?

● **Hint:** You will find it easier to work in eV for Questions 2, 3 and 4 .

The electronvolt
Remember 1 eV is the KE gained by an electron accelerated through a p.d. of 1 V.

$$1\ \text{eV} = 1.6 \times 10^{-19}\ \text{J}$$

Quantum explanation of atomic spectra

We can now put our quantum models of the atom and light together to explain the formation of atomic spectra. To summarize what we know so far:

- Atomic electrons can only exist in certain discrete energy levels.
- Light is made up of photons.
- When electrons lose energy they give out light.
- When light is absorbed by an atom it gives energy to the electrons.

We can therefore deduce that when an electron changes from a high energy level to a low one, a photon of light is emitted. Since the electron can only exist in discrete energy levels, there are a limited number of possible changes that can take place; this gives rise to the characteristic line spectra that we have observed. Each element has a different set of lines in its spectrum because each element has different electron energy levels. To make this clear we can consider a simple atom with electrons in the four energy levels shown in Figure 7.14.

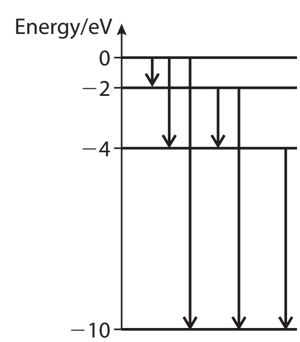


Figure 7.14

As you can see in the diagram there are six possible ways that an electron can change from a high energy to a lower one. Each transition will give rise to a photon of different energy and hence a different line in the spectrum. To calculate the photon frequency we use the formula

$$\text{Change in energy } \Delta E = hf$$

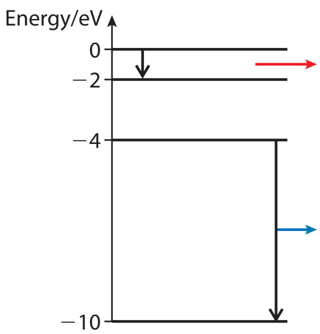


Figure 7.15

So the bigger energy changes will give lines on the blue end of the spectrum and low energies on the red end.

Example

A change from the −4 eV to the −10 eV level will result in a change of 6 eV. This is $6 \times 1.6 \times 10^{-19} = 9.6 \times 10^{-19}$ J

This will give rise to a photon of frequency given by

$$\Delta E = hf$$

Rearranging gives $f = \dfrac{\Delta E}{h} = \dfrac{9.6 \times 10^{-19}}{6.63 \times 10^{-34}} = 1.45 \times 10^{15}$ Hz

This is a wavelength of 207 nm which is UV.

Ionization

Ionization occurs when the electrons are completely removed from an atom, leaving a charged atom called an ion. This can happen if the atom absorbs a high energy photon or the electron could be 'knocked off' by a fast moving particle like an alpha. These interactions are quite different – when a photon interacts with an atom it is absorbed but when an alpha interacts it 'bounces off'.

Comparing energies in eV

- The average KE of an atom in air at 20 °C is about 0.02 eV
  a red light photon is 1.75 eV
  a blue light photon is 3.1 eV.
- The energy released by one molecule in a chemical reaction is typically 50 eV.
- The energy released by one atom of fuel in a nuclear reaction is 200 MeV.

Absorption of light

A photon of light can only be absorbed by an atom if it has exactly the right amount of energy to excite an electron from one energy to another. If light containing all wavelengths (white light) is passed through a gas then the photons with the right energy to excite electrons will be absorbed. The spectrum of the light that comes out will have lines missing. This is called an *absorption spectrum* and is further evidence for the existence of electron energy levels.

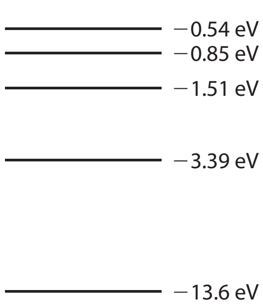


Figure 7.16

Exercises

Use the energy level diagram of Figure 7.16 to answer the following questions.

5 How many possible energy transitions are there in this atom?

6 Calculate the maximum energy (in eV) that could be released and the frequency of the photon emitted.

7 Calculate the minimum energy that could be released and the frequency of the associated photon.

8 How much energy would be required to completely remove an electron that was in the lowest energy level? Calculate the frequency of the photon that would have enough energy to do this.

7.3 The wave nature of matter

Assessment statements

13.1.5 Describe the de Broglie hypothesis and the concept of matter waves.
13.1.6 Outline an experiment to verify the de Broglie hypothesis.
13.1.7 Solve problems involving matter waves.
13.1.13 Outline the Heisenberg uncertainty principle with regard to position–momentum and time–energy.

The electron gun

Imagine you have five identical boxes and each one contains one of the following: a large steel ball, a glass ball, air, sand, or a cat. You have to find out what is inside the boxes without opening them. One way of doing this is to fire a bullet at each. Here are the results:

1. Shattering sound, contents – glass ball
2. Bounces back, contents – steel ball
3. Passes straight through, contents – air
4. Doesn't pass through, contents – sand

Well, the cat was lucky so we won't try shooting at the last box.

This idea is used a lot by physicists to find out what matter contains, but the projectile must be something much smaller than a bullet. In the Geiger and Marsden experiment discussed earlier, the projectile was an alpha particle. Electrons can also be used in this way but first they need to be accelerated, and to do this we use an electron gun, as shown in Figure 7.17.

This is an example of how scientists use experimental evidence to build models of things they can't see. Note that in these examples you don't just need a bullet, you also need a detector. What if you didn't notice the sound of breaking glass? You would have to have some idea that this was going to happen or you might not have been listening.

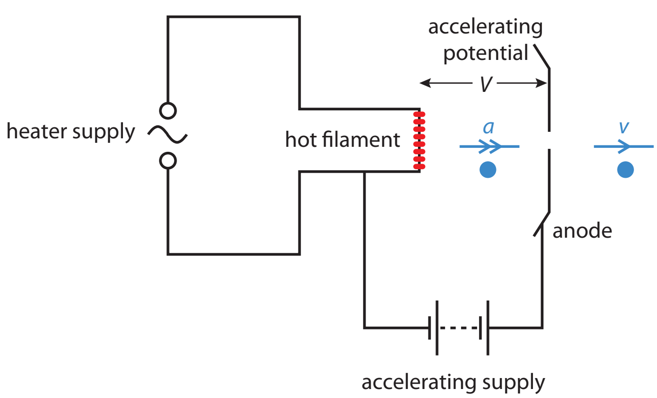


Figure 7.17 A simple electron gun.

Thermionic emission
In this case atomic electrons gain enough energy from thermal excitation to leave the surface of the metal.

The filament, made hot by an AC supply, liberates electrons which are accelerated towards the anode by the accelerating p.d. Electrons travelling in the direction of the hole in the anode pass through and continue with constant velocity.
We can calculate the speed of the electrons, using the law of conservation of energy.

Loss of electrical PE = gain in KE

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

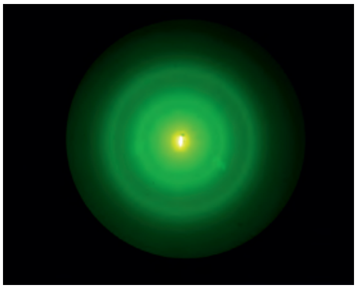
$$v = \sqrt{\frac{2Ve}{m}}$$

KE in eV
The KE of the electrons in eV is numerically the same as the accelerating p.d.
If p.d. = 200 V
KE = 200 eV

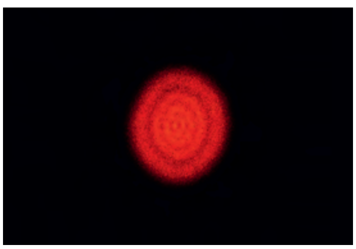
Detecting electrons

You can't see electrons directly, so you need some sort of detector to find out where they go. When electrons collide with certain atoms they give the atomic electrons energy (we say the electrons are 'excited'). When the atomic electrons go back down to a lower energy level, they give out light. This is called *phosphorescence* and can be used to see where the electrons land. Zinc sulphide is one such substance; it is used to coat glass screens so that light is emitted where the electrons collide with the screen. This is how the older types of TV screens work.

The aurora borealis is caused when charged particles from the Sun excite atoms in the atmosphere.



Diffraction of electrons by a graphite film.



Diffraction of light by a small circular aperture.

Electron diffraction

If a beam of electrons is passed through a thin film of graphite, an interesting pattern is observed when they hit a phosphor screen. The pattern looks very much like the diffraction pattern caused when light is diffracted by a small circular aperture. Perhaps the electrons are being diffracted by the atoms in the graphite. Assuming this to be true, we can calculate the wavelength of the wave that has been diffracted. But diffraction is a wave-like property, so what is it about an electron that behaves like a wave?

The de Broglie hypothesis

In 1924 Louis de Broglie proposed that 'all matter has a wave-like nature' and that the wavelength of that wave could be found using the equation:

$$\lambda = \frac{h}{p}$$

where p is the momentum of the particle.

Using this equation we can calculate the momentum of the electrons used in the electron diffraction experiment.

If the accelerating p.d. is 500 V then the velocity of an accelerated electron is

$$v = \sqrt{\frac{2 \times 500 \times 1.6 \times 10^{-19}}{9.1 \times 10^{-31}}} = 1.3 \times 10^{7}\,\mathrm{m\,s^{-1}}$$

Momentum $= mv = 1.2 \times 10^{-23}$ Ns
So the wavelength, $\lambda = 5.5 \times 10^{-11}$ m

This is the correct size to give a wavelength that would cause the observed diffraction pattern. But what is it about the electron that has wave-like properties?

Probability waves

The de Broglie hypothesis makes us think about the nature of particles like electrons in a very different way. Classically we think of an electron as a very small ball. Small balls are very simple to model. If we know where a ball is now and what its momentum is, we can use the *suvat* equations to work out exactly where it will be in the future. According to the de Broglie hypothesis, matter isn't like that; the position of a particle is governed by a wave function, not the *suvat* equations; this wave function gives the probability of finding the particle. When the particle passes through a small opening, the wave diffracts, giving the particle the possibility to be in places we wouldn't expect. This is only evident for very small particles because large particles cannot pass through anything small enough for diffraction to take place. (Remember diffraction is only noticeable when the opening is about the same size as the wavelength.)

Consider electron A on a crash course for electron B as shown in Figure 7.18. Classically we can calculate exactly when they collide. We know when this takes place because B will start to move.

What the de Broglie hypothesis says is that the position of A is given by a probability distribution that is defined by the wave travelling in the same direction. As the probability wave passes B the probability of an interaction becomes higher and higher, until at some moment B will move. We can't say when exactly this will happen, just that it happens some time as the wave passes B. So the electron isn't a little ball at all – it is simply a region of space that has certain properties (charge and mass). These properties define the way that this region of space can interact with other regions and the probability of an interaction is given by a wave equation.

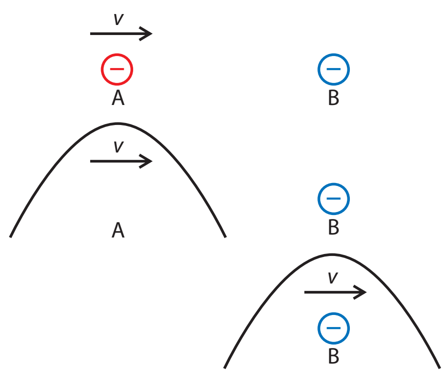


Figure 7.18 Electron A approaches electron B.

The Davidson-Germer experiment
In this experiment a beam of electrons was reflected off a nickel crystal. The angle of the maximum intensity of reflected electrons can be explained in terms of the constructive interference of de Broglie waves reflected off different layers of atoms. This supports the de Broglie hypothesis.

Explaining diffraction

Using the de Broglie hypothesis, we can now attempt to explain electron diffraction. An electron is travelling toward a very narrow slit. As it travels, its wave function precedes it. The wave function will pass through the slit, mapping out the probable positions of the electron. This wave function is diffracted by the slit, which makes it possible for the electron to appear in the maxima of the pattern. If lots of electrons pass through the slit then they will all land somewhere in the maxima of the pattern, forming the electron diffraction pattern seen in the previous photo.

Exercises

9 An electron is accelerated by a p.d. of 100 V. Calculate its
(a) KE in eV
(b) KE in joules
(c) de Broglie wavelength.

10 Calculate the de Broglie wavelength for a car of mass 1000 kg travelling at 15 m s^{-1}. Why won't you ever see a car diffract?

Heisenberg's uncertainty principle

To find the path of a particle we need to know its position and momentum. Let us try to limit the position of an electron in a parallel beam of electrons all with equal momentum. We can do this by passing it through a narrow slit as in Figure 7.19. The uncertainty in its position Δx as it passes through the slit is therefore equal to the slit width. If the slit is wide, then the electron will pass straight through without diffracting, so we know its future direction and hence its momentum, but the uncertainty in position is large. To reduce this uncertainty, we can reduce the slit width, but if we do that it will diffract. This gives an increased number of possibilities for its path, so the uncertainty in momentum increases. This no-win situation is called the *Heisenberg uncertainty principle* and can be expressed by the equation:

$$\Delta p \Delta x \geqslant \frac{h}{4\pi}$$

A similar relationship exists when you try to define energy and time.

$$\Delta E \Delta t \geqslant \frac{h}{4\pi}$$

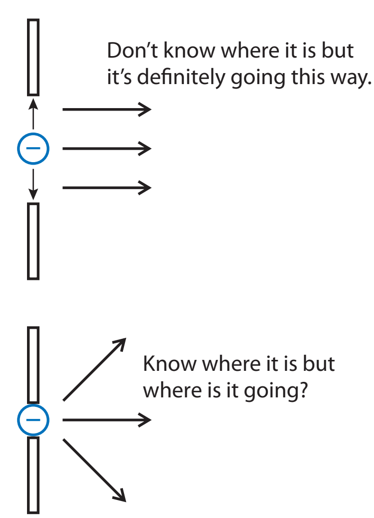


Figure 7.19 We can't know both momentum and position accurately.

It is important to realise that the uncertainty principle is not just about measurement – it is about the way things are.

7.4 Quantum models of the atom

Assessment statements

13.1.11 Explain the origin of atomic energy levels in terms of the 'electron in a box' model.

13.1.12 Outline the Schrödinger model of the hydrogen atom.

Returning to the atomic model, we know that atomic electrons can only have discrete energy levels, which are different for each element. Treating atomic electrons as waves gives us an insight into the reason why electrons can only have certain discrete energy levels.

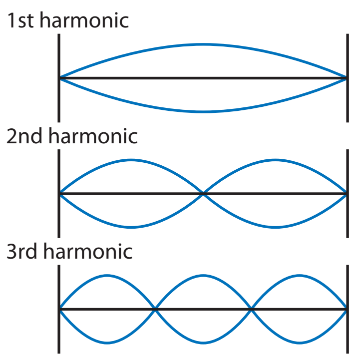


Figure 7.20

The electron in a box

The probability wave that defines the position of a free electron can be thought of as a wave travelling along a string. As a peak moves down the string, the most likely position of the electron moves with it. An atomic electron is not free to move outside the atom, and to model this we can think of a wave trapped in a short length of string clamped at both ends (like a guitar string). When we trap a wave in a string it reflects backwards and forwards, creating a standing wave. This standing wave must have nodes at either end, so can only have certain frequencies, the harmonics, as shown in Figure 7.20. This is very similar to the situation with an atomic electron that can only have certain discrete energies. By treating an atomic electron as a probability wave trapped in a box, we can develop a quantum model of the atom.

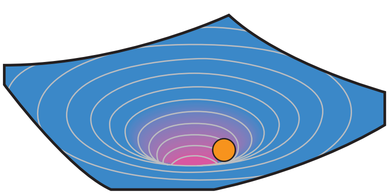


Figure 7.21 An electron trapped in a potential well.

To pull an electron out of an atom you would need to exert a force to oppose the electric attraction between the electron and the nucleus. We can think of the electron as a ball in a potential well, as shown in Figure 7.21; the shape of the sides of the well is determined by the electric force. This is rather complicated, so to make life simpler, we will consider the potential well to be a square box.

The wavelength of the electron's probability wave is governed by the size of the box, the longest possible wavelength being $2L$ as shown in Figure 7.23.

The de Broglie wavelength, $\lambda = \dfrac{h}{p} = \dfrac{h}{mv}$

so $\qquad v = \dfrac{h}{m\lambda}$

The KE of the electron $= \frac{1}{2}mv^2$

Substituting for v

gives KE $= \frac{1}{2}m \times \dfrac{h^2}{m^2\lambda^2} = \dfrac{h^2}{2\lambda^2 m}$

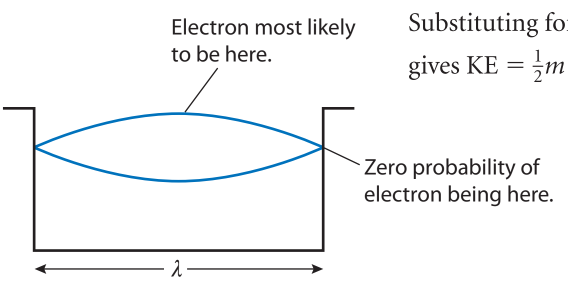


Figure 7.22 The probability wave confined to a square box.

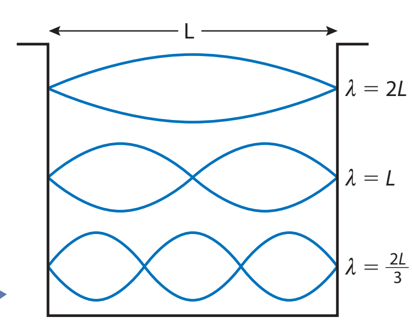


Figure 7.23

We know that the wavelength can have many values, as in Figure 7.23:

$$2L, \frac{2L}{2}, \frac{2L}{3}, \frac{2L}{4}$$

All possible waves can be defined by the equation $\lambda = \frac{2L}{n}$
where n is an integer (1,2,3....)

Substituting into the equation for KE gives

$$\text{KE} = \frac{n^2 h^2}{8mL^2}$$

So by treating the electron as a wave trapped in a box, we find that the energy can only be certain discrete values. From this we can predict the energy levels of the hydrogen atom and hence the possible energy level changes that give rise to the atomic spectra. Unfortunately this does not give the right result, as the model is a bit too simple. However, we are on the right track.

Schrödinger's model

Schrödinger realised that the probability that the electron is in a particular position in the atom is defined by the square of a wave function, but that wave function was not as simple as the sine wave used previously. The wave function is called *Schrödinger's equation* and is denoted by the Greek letter ψ. The probability of finding the electron is therefore ψ^2. By solving the equation for this wave trapped in the potential well of the atom, Schrödinger calculated the correct values for the possible electron energies. His model also predicted that some energy transitions were more likely, explaining why some spectral lines are brighter than others.

If we plot the probability function for the different energies of the electron in hydrogen we see that for a given energy there are some places that the electron is more likely to be than others, but the electron is not stationary – it is moving about very quickly. So if we could watch the electron, it would appear like a cloud, even though there is only one electron. For the hydrogen atom, some of these clouds form spheres of different radius, the larger energy clouds having bigger radius.

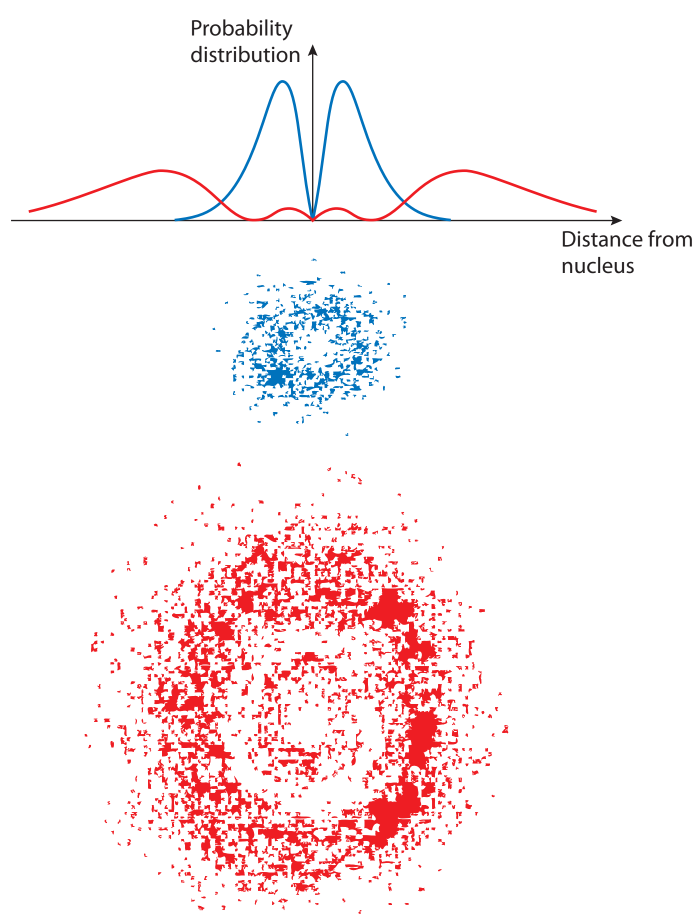


Figure 7.24 The probability densities and electron clouds for two of the possible electron energies for the hydrogen atom.

7.5 Nuclear structure

Assessment statements

7.1.5 Explain the terms nuclide, isotope and nucleon.

7.1.6 Define nucleon number A, proton number Z and neutron number N

7.1.7 Describe the interactions in a nucleus.

7.3.3 Define the term unified atomic mass unit.

7.3.4 Apply the Einstein mass–energy equivalence relationship.

7.3.5 Define the concepts of mass defect, binding energy and binding energy per nucleon.

7.3.6 Draw and annotate a graph showing the variation with nucleon number of the binding energy per nucleon.

7.3.7 Solve problems involving mass defect and binding energy.

13.2.2 Describe how the masses of nuclei may be determined using a Bainbridge mass spectrometer.

13.2.1 Explain how the radii of nuclei may be estimated from charged particle scattering experiments.

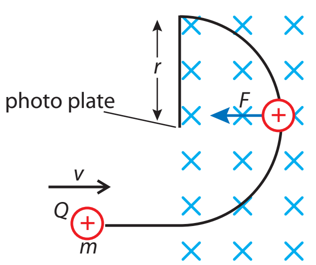


Figure 7.25 Deflection chamber of a mass spectrometer.

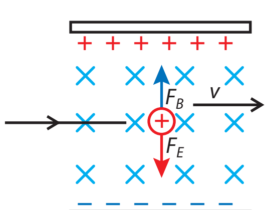


Figure 7.26 Velocity selector of a mass spectrometer.

Mass of the nucleus

The mass of a nucleus can be measured using a mass spectrometer. The principle behind the operation of a mass spectrometer is that if nuclei of the same charge and velocity are projected at right angles to a region of uniform magnetic field, the radius of the resulting circular path will be proportional to their mass. Figure 7.25 shows this situation.

The centripetal force is provided by the magnetic force, so we can write the equation:

$$\frac{mv^2}{r} = BQv$$

Rearranging this gives $m = \frac{BQr}{v}$

So the mass is proportional to the radius, provided the velocity and charge are the same.

In practice, it is difficult to remove all the electrons from an atom, leaving just the nucleus, but it is enough to just remove one. This leaves the atom with a charge of $+e$, the extra mass of the electrons making little difference. If we have a gas of ionized atoms then we can increase their velocity by heating. However, as we know from thermal physics, the atoms of a gas do not all have the same velocity. To select atoms with the same velocity, a combination of electric and magnetic fields is used (as shown in Figure 7.26).

An ion entering the velocity selector will experience two forces:

magnetic force $F_B = BQv$ (upwards) and

electric force $F_E = EQ$ (downwards).

These forces will be balanced if $F_E = F_B$, and in this case the particle will continue with a constant velocity.

For the forces to be balanced, $BQv = EQ$

This is only true when $v = \frac{E}{B}$ so all the ions that continue in a straight line must have velocity equal to $\frac{E}{B}$.

To detect the ions, a photographic plate is positioned as in Figure 7.25. When ions land on the plate, they change the chemical structure of the plate, which results in a small black dot when the plate is developed. Counting the number of dots enables the number of ions to be determined.

Nuclear masses are measured in unified mass units (u).
Proton mass = 1.0078 u.
Neutron mass = 1.0086 u.

1 u is the mass of $\frac{1}{12}$ of the nucleus of a carbon-12 isotope.

Exercises

Assume all ions have a charge of $+e$.

11 Considering the mass spectrometer in Figure 7.27, find

(a) the velocity of the ions coming through the velocity selector

(b) the mass of an ion with a deflection radius of 6 cm.

12 An ion of mass 5.1×10^{-26} kg passes into the deflection chamber of the mass spectrometer in Question 11. What is its radius of deflection?

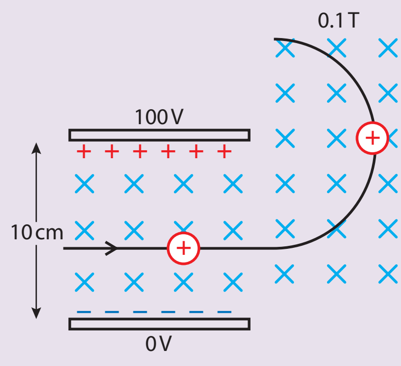


Figure 7.27

By measuring the masses of all known nuclei it was discovered that they all have a mass that is a multiple of the same number. This led to the idea that the nucleus is made up of smaller particles. In the same way, if you measured the mass of many boxes of apples and found that the mass of every box was a multiple of 100 g, you might conclude that the mass of each apple was 100 g.

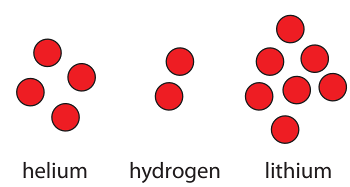


Figure 7.28 Some different sized nuclei.

Charge of the nucleus

It is also possible to measure the charge of all known nuclei. Each nucleus is found to have a charge that is a multiple of the charge on the electron, but positive. It would be reasonable to think that each of the particles in the nucleus must have a positive charge, but that makes the charge too big. It is therefore suggested that there are two types of particle in the nucleus, protons with positive charge and neutrons with no charge.

Masses and charges
Proton mass = 1.672×10^{-27} kg
Neutron mass = 1.675×10^{-27} kg

Proton charge = $+1.6 \times 10^{-19}$ C
Neutron charge = 0

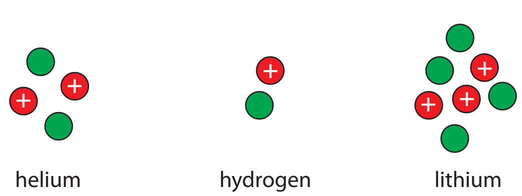
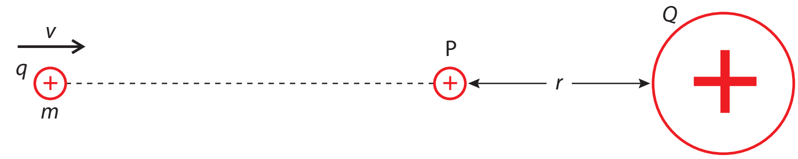


Figure 7.29 Nuclei showing protons and neutrons.

Size of the nucleus

The size of the nucleus can be determined by conducting a similar experiment to the alpha scattering experiment of Geiger and Marsden. The alpha particles that come straight back off the gold foil must have approached a nucleus head-on following a path as shown in Figure 7.30.

Figure 7.30



Applying the law of conservation of energy to this problem, we can deduce that at point P, where the alpha particle stops, the original KE has been converted to electrical PE.

$$\tfrac{1}{2}mv^2 = \frac{kQq}{r}$$

where Q = the charge of the nucleus ($+Ze$)
and q = the charge of the alpha ($+2e$)

When Geiger and Marsden did their gold foil experiment they didn't have alphas with enough energy to get close enough to the gold nucleus to give a very accurate value for its radius. They did however get quite close to aluminium nuclei.

The KE of the alpha can be calculated from the change in the mass of a nucleus when it is emitted (more about this later), so knowing this, the distance r can be calculated.

To determine the size of the nucleus, faster and faster alphas are sent towards the nucleus until they no longer come back. The fastest ones that return have got as close to the nucleus as possible.

This is just an estimate, especially since (as is the case for all particles) the position of the particles that make up the nucleus is determined by a probability function. This will make the definition of the edge of the nucleus rather fuzzy.

Exercise

13 It is found that alpha particles with KE 7.7 MeV bounce back off an aluminium target. If the charge of an aluminium nucleus is $+2.1 \times 10^{-18}$ C, calculate:
(a) the velocity of the alpha particles (they have a mass of 6.7×10^{-27} kg)
(b) the distance of closest approach to the nucleus.

Some quantities and terms related to the nucleus

- **Nucleon**

The name given to the particles of the nucleus (proton, neutron).

- **Nuclide**

A combination of protons and neutrons that form a nucleus.

- **Isotopes**

Nuclei with the same number of protons but different numbers of neutrons (as shown in Figure 7.31).

- **Nucleon number** (*A*)

The number of protons plus neutrons in the nucleus (e.g. lithium-7). This will be different for different isotopes of the same element.

- **Proton number** (Z)

The number of protons in the nucleus (e.g. for lithium this is 3). This is always the same for a given element.

- **Neutron number** (N)

The number of neutrons in a nucleus.

- **Symbol for a nucleus**

A nuclide can be represented by a symbol that gives details of its constituents,

e.g. $^{7}_{3}\mathrm{Li}$ 7 – nucleon number
3 – proton number

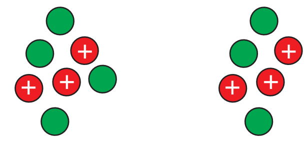


Figure 7.31 Two isotopes of lithium.

Exercises

14 How many protons and neutrons are there in the following nuclei?

(a) $^{35}_{17}\mathrm{Cl}$

(b) $^{58}_{28}\mathrm{Ni}$

(c) $^{204}_{82}\mathrm{Pb}$

15 Calculate the charge in coulombs and mass in kg of a $^{54}_{26}\mathrm{Fe}$ nucleus.

16 An isotope of uranium (U) has 92 protons and 143 neutrons. Write the nuclear symbol for this isotope.

17 Describe the structure of another isotope of uranium, having the symbol $^{238}_{92}\mathrm{U}$.

The nuclear force (strong force)

If a metal wire is heated it will readily emit electrons. However, you would have to heat it to temperatures approaching the temperature of the Sun before any neutrons or protons broke free. The fact that it requires a lot of energy to pull the nucleus apart implies that the force holding the nucleons together is a very strong force. It is therefore not surprising that when particles do come out of the nucleus, they have very high energy. We know that protons are positive, so they must repel each other. However, the nuclear force is so much bigger than the electric force that the nucleons are not pushed apart.

Although the nuclear force is strong, different nuclei do not attract each other. This means that the force must be very short range, unlike the electric force that extends forever.

Since protons repel each other, we might expect that the nuclear force would be stronger between protons than between neutrons. If this were the case, then nuclides with more neutrons would be pulled together more tightly than those with a lot of protons. But since all nuclei have the same density, this is not the case.

So the nuclear force is very strong, short range and the same for all nucleons.

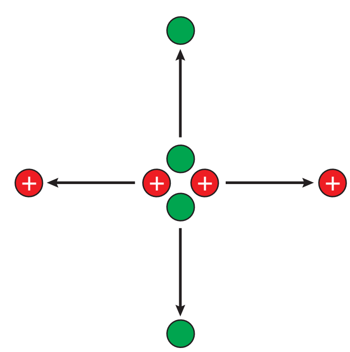


Figure 7.32 Work must be done to pull a nucleus apart; if a nucleus is put together then energy must be released.

Binding energy

The binding energy is defined as the amount of work required to pull apart the constituents of a nucleus.

If work is done, then energy must have been transferred to the nucleons. However, they aren't moving so haven't any KE, and they aren't in a field, so don't have PE.

Where has the energy gone?

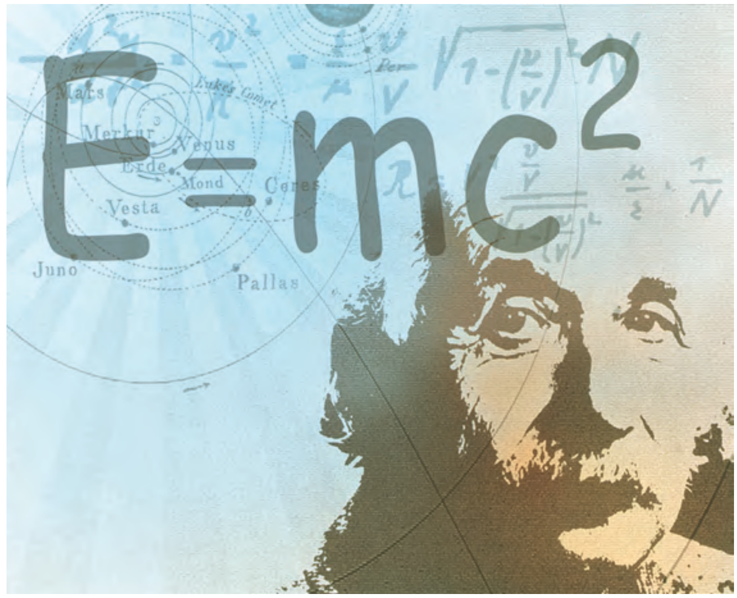
Einstein solved this problem with his famous equation

$$E = mc^2$$

We measure nuclear energies in MeV. 1 MeV is the KE an electron would gain if accelerated through a p.d. of 1 000 000 V.

The energy has been converted into mass. This would imply that the mass of the particles when they are apart is greater than when they are together. If a nucleus has a large binding energy then it will require a lot of work to pull it apart – we say it is stable.

$E = mc^2$ is probably the most famous formula in the world – and Albert Einstein, definitely the most famous physicist.



Mass equivalents
1 kg = 9×10^{16} J
1 u = 931.5 MeV

The binding energy curve

It is possible to calculate the binding energy (BE) of a nucleus by finding the difference between the mass of the nucleus and the mass of the parts – this is called the *mass defect*. We can then plot a graph of BE per nucleon against nucleon number for all known nuclei. The result is shown below.

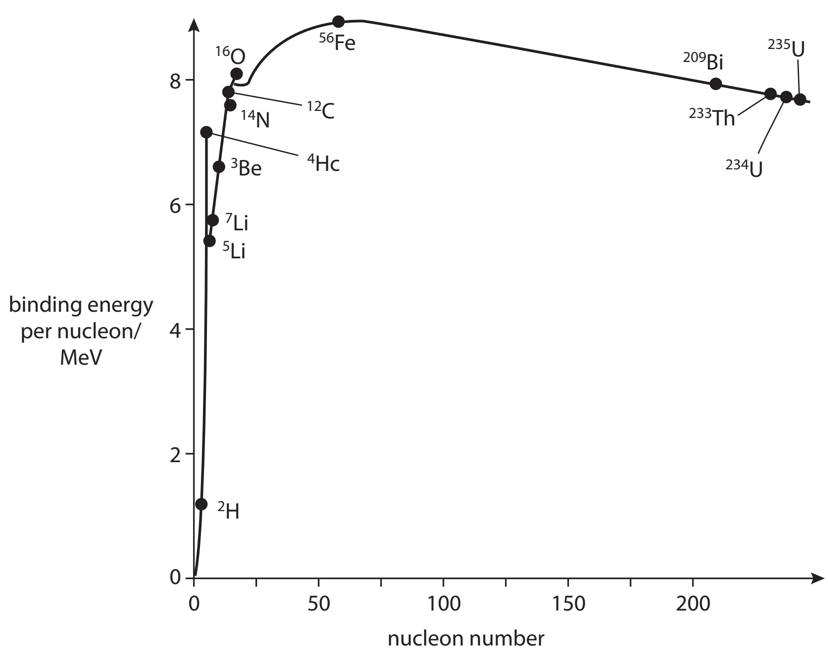


Figure 7.33 Graph of BE/nucleon vs nucleon number.

From this chart we can see that some nuclei are more stable than others. Iron is in fact the most stable, which is why there is so much of it. In nuclei where there are a lot of protons the electric repulsion tends to push them apart. This means that large nuclei are less stable, and there are no nuclei that have more than about 300 nucleons.

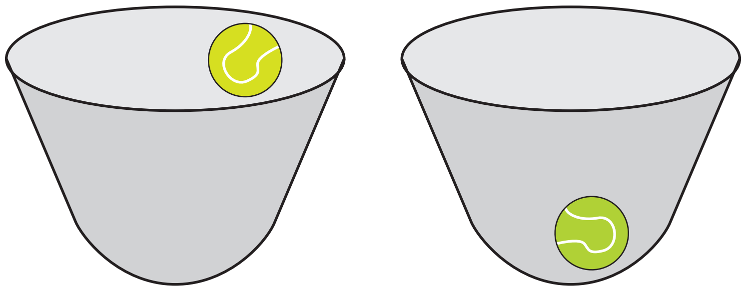


Figure 7.34 To lift a ball out of the bottom of a bowl we need to do work. This is just like the BE of the nucleus. When the ball is in the bowl, it does not have this energy – it is what was lost when it rolled to the bottom.

We have found that all physical systems will, if possible, reach a position of lowest possible energy. So, if possible, a nucleus will change into one with lower energy – this means that a nucleus will change to one with more BE. Remember, BE is released when the nucleus is formed, so changing to a higher BE means energy is released.

As we can see from the section of the graph shown in Figure 7.35, ^{233}Th has a higher BE than ^{235}U. It would therefore be energetically favourable for ^{235}U to change into ^{233}Th. However this may not be possible.

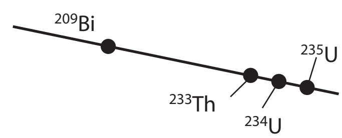


Figure 7.35 A section of the BE curve.

Nuclear masses			
Z	Symbol	*A*	Mass (u)
1	H	1	1.0078
1	D	2	2.0141
2	He	3	3.0160
3	Li	6	6.0151
4	Be	9	9.0122
7	N	14	14.0031
17	Cl	35	34.9689
26	Fe	54	53.9396
28	Ni	58	57.9353
36	Kr	78	77.9204
38	Sr	84	83.9134
56	Ba	130	129.9063
82	Pb	204	203.9730
86	Rn	211	210.9906
88	Ra	223	223.0185
92	U	233	233.0396

The table gives a selection of nuclear masses measured in u. With these values it is possible to calculate the binding energy of the nucleus.

Database
For a fully searchable database of all nuclides , visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 7.3. Click on the element in the Periodic Table to find nuclear data.

Worked example

Calculate the binding energy of iron (Fe).

Nucleon	Mass (u)
Proton	1.00782
Neutron	1.00866

Solution

From the table we can see that the iron nucleus is made of 26 protons and $(54 - 26) = 28$ neutrons.

The total mass of the nucleons that make up iron is therefore:

$26 \times$ mass of a proton $+ 28 \times$ mass of a neutron $= 54.4458$ u

But from the table, the mass of iron is 53.9396 u.

The difference between these two values is 0.5062 u.

But 1 u = 931.5 MeV, so this is equivalent to an energy of 0.5062×931.5 MeV

BE = 471.5 MeV

Since there are 54 nucleons in iron then the binding energy per nucleon is

$\frac{471.5}{54} = \mathbf{8.7\ MeV/nucleon}$

Exercises

18 Find uranium in the table of nuclear mass.

- **(a)** How many protons and neutrons does uranium have?
- **(b)** Calculate the total mass of the protons and neutrons that make uranium.
- **(c)** Calculate the difference between the mass of uranium and its constituents (the mass defect).
- **(d)** What is the binding energy of uranium in MeV?
- **(e)** What is the BE per nucleon of uranium?

19 Enter the data from the table into a spreadsheet. Add formulae to the spreadsheet to calculate the binding energy per nucleon for all the nuclei and plot a graph of BE/nucleon against nucleon number.

Remember: 1u is equivalent to 931.5 MeV.

7.6 Radioactive decay

Assessment statements

7.2.1 Describe the phenomenon of natural radioactive decay.
7.2.2 Describe the properties of α- and β-particles and γ-radiation.
7.2.3 Describe the ionizing properties of α- and β-particles and γ-radiation.
7.2.4 Outline the biological effects of ionizing radiation.
7.2.5 Explain why some nuclei are stable while others are unstable.

13.2.3 Describe one piece of evidence for the existence of nuclear energy levels.
13.2.4 Describe β^+ decay, including the existence of the neutrino.

The nucleus can lose energy by emitting radiation. There are three types of ionizing radiation: alpha, beta and gamma. Alpha and beta emissions result in a change in the number of protons and neutrons. Gamma is a form of electromagnetic radiation, similar to X-rays. When a nucleus changes in this way it is said to *decay*.

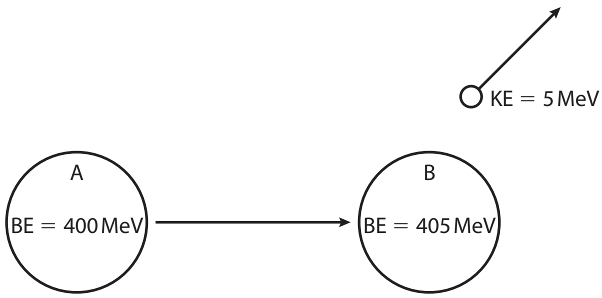


Figure 7.36 A nucleus can emit radiation only if the result is an increase in binding energy.

If nucleus A were made from its constituent protons and neutrons, 400 MeV of energy would be released, but making B would release 405 MeV. Therefore if A changes to B, 5 MeV of energy must be released. This excess energy is given to the particle emitted.

Alpha radiation (α)

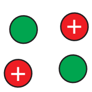
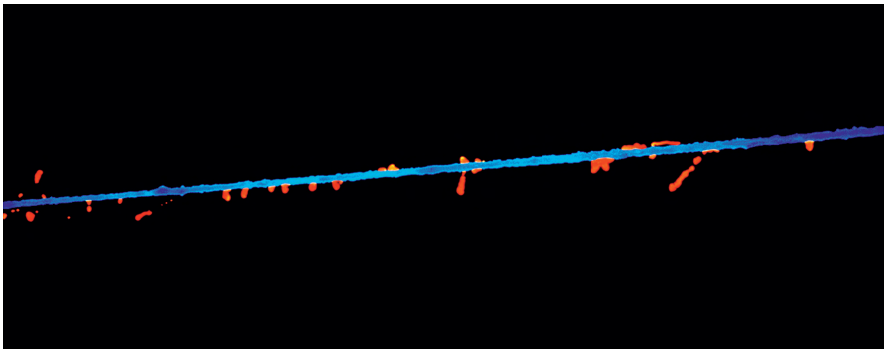


Figure 7.37 An alpha particle is a helium nucleus (2 protons + 2 neutrons).

Alpha particles have energies of about 5 MeV. To knock an electron out of an atom requires about 10 eV, so alpha particles can ionize a lot of atoms before they lose all their KE. This property makes them very easy to detect using a Geiger-Muller tube, photographic paper or cloud chamber. It also makes them harmful since ionizing atoms of human tissue causes damage to the cells similar to burning. Due to their high reactivity and ionization, alpha particles have a range of only a few centimetres in air and cannot penetrate paper.



Cloud chamber track of an alpha particle. These are like mini vapour trails. The red lines are some of the electrons that the alpha has knocked off atoms in the chamber.

Effect on nucleus

When a nucleus emits an alpha particle it loses 2 protons and 2 neutrons. The reaction can be represented by a nuclear equation. For example, radium decays into radon when it emits an alpha particle.

$$^{226}_{88}\text{Ra} \longrightarrow {}^{222}_{86}\text{Rn} + {}^{4}_{2}\text{He}$$

4 nucleons are emitted, so the nucleon number is reduced by 4.
2 protons are emitted, so the proton number is reduced by 2.

Energy released

When radium changes to radon the BE is increased. This leads to a drop in total mass, this mass having been converted to energy.

Mass of radium > (mass of radon + mass of alpha)

Energy released = $\{\text{mass}_{Ra} - (\text{mass}_{Rn} + \text{mass}_{alpha})\}c^2$

$\text{mass}_{Ra} = 226.0254\,\text{u}$

$\text{mass}_{Rn} = 222.0176\,\text{u}$

$\text{mass}_{He} = 4.002602\,\text{u}$

Change in mass = 0.005198 u

This is equivalent to energy of 0.005198 × 931.5 MeV

Energy released = 4.84 MeV

Alpha energy

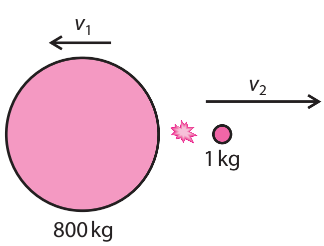
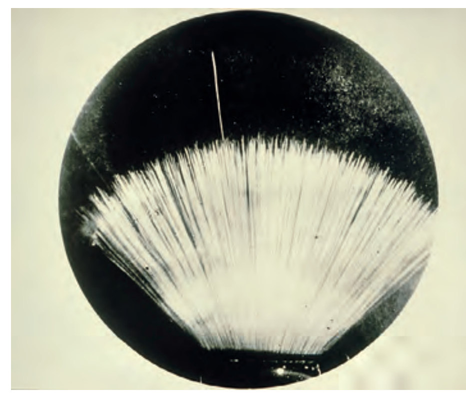


Figure 7.38

When an alpha particle is ejected from a nucleus, it is like an explosion where a small ball flies apart from a big one, as shown in Figure 7.38. An amount of energy is released during the explosion, which gives the balls the KE they need to fly apart.



Alpha tracks in a cloud chamber. Note they are all the same length except for one.

Applying the principle of conservation of momentum:

momentum before = momentum after

$0 = 1 \times v_2 - 800 \times v_1$

Therefore $v_1 = \dfrac{v_2}{800}$

The velocity of the big ball is therefore much smaller than the small one, which means that the small ball gets almost all the kinetic energy. The energy available when alpha particles come from a certain nuclear reaction is constant; it can be calculated from the change in mass. We can see this if we look at the cloud chamber tracks in the photo. As the particles pass through the vapour they ionize vapour atoms, each ionization leading to a loss in energy. The more KE the particle has, the more ionizations it can make; this leads to a longer track. We can see in the photo that most of the tracks are the same length, which means all the alphas had the same KE, say 5 MeV. However if we look carefully at the photo, we can see there is one track that is longer. From the length of the track we can calculate it has energy of 8 MeV. If we calculate the energy released when the nucleus decays, we find it is 8 MeV, which means that when the nucleus is giving out the alphas with low energy, it is remaining in an excited state. This energy will probably be lost at a later time by the emission of gamma radiation. The nucleus seems to have energy levels like atomic electrons. Figure 7.39 shows the energy levels that could cause the change in this example.

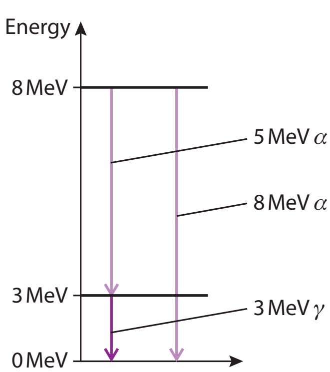


Figure 7.39

Properties of alpha radiation	
Range in air	~5 cm
Penetration	stopped by paper
Ionizing ability	very high
Detection	GM tube, cloud chamber, photographic paper

Beta minus (β−)

Beta minus particles are electrons. They are exactly the same as the electrons outside the nucleus but they are formed when a neutron changes to a proton. When this happens an antineutrino is also produced.

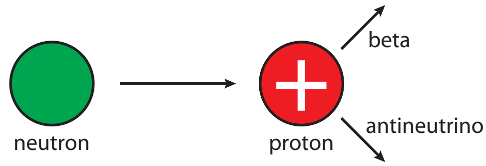
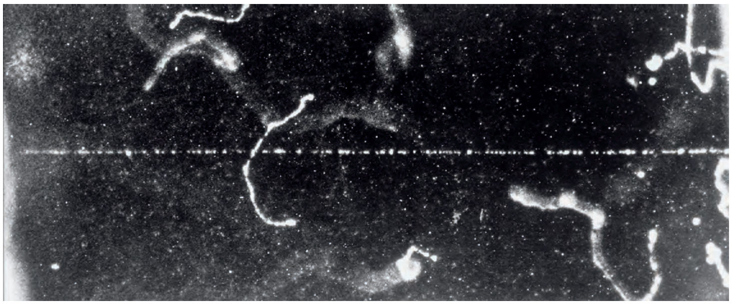


Figure 7.40 Beta minus decay.

It appears that a neutron must be made of a proton with an electron stuck to it, but this isn't the case. Electrons cannot exist within the nucleus. One reason is that electrons are not affected by the strong force that holds the nucleons together, and the electric attraction between proton and electron is not strong enough to hold the electrons in place. During beta decay, the neutron changes into the proton and electron, rather like the frog that changes to a prince in the fairy tale.

Beta particles are not as heavy as alphas and although they travel with high speed they are not as effective at knocking electrons off atoms. As a result they are not as ionizing, although they do produce enough ions to be detected by a GM tube, cloud chamber or photographic plate. Since betas are not as reactive or ionizing as alphas, they pass through more matter and have a longer range in air.



A cloud chamber image of a beta track – notice it is much thinner than the alpha tracks.

Properties of beta radiation	
Range in air	~30 cm
Penetration	~1 mm aluminium
Ionizing ability	not very
Detection	GM tube, cloud chamber, photographic paper

Effect on nucleus

When a nucleus emits a beta particle, it loses 1 neutron and gains 1 proton. Carbon-14 decays into nitrogen-14 when it emits a beta particle.

$$^{14}_{6}\mathrm{C} \rightarrow {}^{14}_{7}\mathrm{N} + e^{-} + \bar{\nu}$$

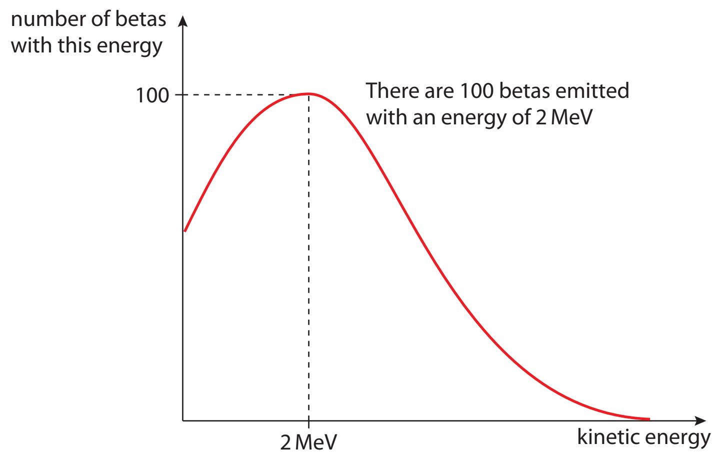
Beta energy

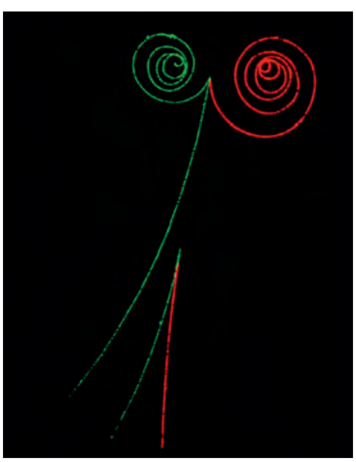
As with alpha decay, the amount of energy available to the beta particles can be calculated from the change of mass that occurs. However, in this case, the energy is shared between the beta and the neutrino. If we measure the KE of the beta particles, we find that they have a range of values, as shown in the beta energy spectrum in Figure 7.41. When beta particles were first discovered, the existence

Antineutrino
An antineutrino is the antiparticle of a neutrino. All particles have antiparticles. You will understand more about antiparticles if you do the particle physics option.

of the neutrino was unknown, so physicists were puzzled as to how the beta could have a range of energies; it just isn't possible if only one particle is ejected from a heavy nucleus. To solve this problem, Wolfgang Pauli suggested that there must be an 'invisible' particle coming out too. The particle was called a *neutrino* (small neutral one) since it had no charge and negligible mass. Neutrinos are not totally undetectable and were eventually discovered some 26 years later. However, they are very unreactive; they can pass through thousands of kilometres of lead without an interaction.

Figure 7.41 This graph shows the spread of beta energy. It isn't really a line graph but a bar chart with thin bars. Notice how many betas have zero KE – this is because they are slowed down by electrostatic attraction to the nucleus.





Electrons (green) and positrons (red) are formed from gamma photons. Notice how the B field (into page) causes them to curve in opposite directions. The bottom pair has more KE than the top pair, so the radius of their path is greater.

Beta-plus ($\beta+$) decay

A beta-plus is a positive electron, or positron. They are emitted from the nucleus when a proton changes to a neutron. When this happens, a neutrino is also produced.

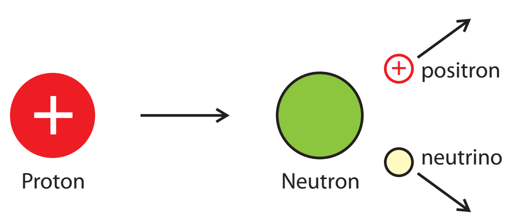


Figure 7.42 Beta-plus decay.

A positron has the same properties as an electron but it is positive (it is the antiparticle of an electron), so beta-plus particles have very similar properties to beta-minus. They penetrate paper, have a range of about 30 centimetres in air, and are weakly ionizing. The beta-plus track in a cloud chamber also looks the same as beta-minus, unless a magnetic field is added. The different charged paths then curve in opposite directions as shown in the photo.

The mass of a neutron is greater than the mass of a proton, so it is possible for a neutron to change to a proton outside the nucleus. $\beta+$ decay is, however, only possible in a nucleus.

Effect on nucleus

When a nucleus emits a $\beta+$ it loses one proton and gains a neutron. An example of a $\beta+$ decay is the decay of sodium into neon.

$$^{22}_{11}\text{Na} \longrightarrow {}^{22}_{10}\text{Ne} + e^{+} + \nu$$

Exercises

20 $^{45}_{20}$Ca (calcium) decays into Sc (scandium) by emitting a β-particle. How many protons and neutrons does Sc have?

21 Cs (caesium) decays into $^{137}_{56}$Ba (barium) by emitting a β-particle. How many protons and neutrons does Cs have?

Antimatter

It is also possible for a proton to change into a neutron and a positive electron plus a neutrino. The positron is in some way the opposite of an electron — it is called an *antiparticle*. The antineutrino $\bar{\nu}$ is also the antiparticle of the neutrino. Every particle has an antiparticle. Atoms made of negative positrons and positive electrons are called *antimatter*.

Gamma radiation (γ)

Gamma radiation is electromagnetic radiation, so when it is emitted there is no change in the particles of the nucleus – they just lose energy. Each time a nucleus decays, a photon is emitted. As we have seen, the energy released from nuclear reactions is very high, so the energy of each gamma photon is also high. The frequency of a photon is related to its energy by the formula $E = hf$. This means that the frequency of gamma photons is very high. Their high energy also means that if they are absorbed by atomic electrons, they give the electrons enough energy to leave the atom. In other words they are ionizing – which means they can be detected with a GM tube, photographic paper or a cloud chamber. As they pass easily through human tissue, gamma rays have many medical applications.

Gamma energy

Gamma photons are often emitted when a nucleus is left in an excited state after emitting another form of radiation e.g. beta.

Consider the following example of beta decay

$$^{12}_{5}\text{B} \rightarrow {}^{12}_{6}\text{C} + \beta^{-} + \bar{\nu}$$

The BE of ^{12}B is 79 MeV and the BE of ^{12}C is 92 MeV

The energy released during this decay is therefore:

$$92 - 79 = 13\,\text{MeV}$$

This can result in a maximum β energy of 13 MeV.

Alternatively, only 10 MeV could be given to the β leaving the nucleus in an excited state. The extra energy could then be released as a gamma photon of energy 3 MeV.

Decay chains

Radioactive nuclei often decay into other nuclei that are also radioactive; these in turn decay, forming what is known as a *decay chain*. An example of a decay chain is the uranium-235 series that starts with plutonium and ends with lead. The table below includes the isotopes at the start of this series.

Table of nuclear masses			
Z	Symbol	*A*	Mass (u)
94	Pu	239	239.052156
92	U	235	235.043923
91	Pa	231	231.035883
90	Th	231	231.036304
90	Th	227	227.027704
89	Ac	227	227.027752

Note:
For the alpha decays in the following exercises, you will also need to know that the mass of an alpha particle is 4.002602 u. But you don't need to include the mass of the beta since it is already taken into account. This is because the masses are actually atomic masses, which include the electrons in a neutral atom. If you add up the electrons, you will discover that there is already an extra one on the left of the equation.

Exercises

22 State whether the following are α- or β-decays.

(a) $^{239}\mathrm{Pu} \rightarrow {}^{235}\mathrm{U}$

(b) $^{235}\mathrm{U} \rightarrow {}^{231}\mathrm{Th}$

(c) $^{231}\mathrm{Th} \rightarrow {}^{231}\mathrm{Pa}$

23 In each of the examples above, use the information in the table to calculate the energy released.

Nuclear radiation and health

Alpha and beta particles have energies measured in MeV. To ionize an atom requires about 10 eV, so each particle can ionize 10^5 atoms before they have run out of energy. When radiation ionizes atoms that are part of a living cell, it can affect the ability of the cell to carry out its function or even cause the cell wall to be ruptured. If a large number of cells that are part of a vital organ are affected then this can lead to death. In minor cases the effect is similar to a burn. The amount of harm that radiation can cause is dependent on the number and energy of the particles. When a gamma photon is absorbed, the whole photon is absorbed so one photon can ionize only one atom. However, the emitted electron has so much energy that it can ionize further atoms, leading to damage similar to that caused by alpha and beta.

A radiation burn caused during radiotherapy for cancer.

Very high dose

- Can affect the central nervous system, leading to loss of coordination and death within two or three days.

Medium dose

- Can damage the stomach and intestine, resulting in sickness and diarrhoea and possibly death within weeks.

Low dose

- Loss of hair, bleeding and diarrhoea.

Safe dose

- All ionizing radiation is potentially harmful so there is no point below which it becomes totally safe. However, at very low levels the risk is small and can be outweighed by the benefits gained when, for example, an X-ray is taken of a broken leg.

Long term

- There is some evidence that after exposure to radiation, the probability of getting cancer or having a child with a genetic mutation increases.

A lot of the data about the effect of radiation comes from studying the victims of the atom bombs dropped on Japan in the Second World War. Is it ethically correct to use this data?

Cancer

Rapidly dividing cancer cells are very susceptible to the effects of radiation and are more easily killed than normal cells. In radiotherapy, nuclear radiation is used to cure cancer by killing the cancerous cells.

Protection against radiation

There are two ways that we can reduce the effect of nuclear radiation: distance and shielding. Alpha and beta radiation have a very short range in air, so will not be dangerous a few metres from the source. The number of gamma photons decreases proportional to $\frac{1}{r^2}$ (where r is the distance from the source), so the further away you are, the safer you will be. Although alpha is the most ionizing radiation, it can be stopped by a sheet of paper (although this means that alpha is the most harmful if ingested). Beta and gamma are more penetrating, so need a thick lead shield to provide protection.

A suit designed to protect the wearer from radiation.

7.7 Half-life

Assessment statements

7.2.6 State that radioactive decay is a random and spontaneous process and that the rate of decay decreases exponentially with time.
7.2.7 Define the term *radioactive half-life*.
7.2.8 Determine the half-life of a nuclide from a decay curve.
7.2.9 Solve radioactive decay problems involving integral numbers of half-lives.

13.2.5 State the radioactive decay law as an exponential function and define the decay constant.
13.2.6 Derive the relationship between decay constant and half-life.
13.2.7 Outline methods for measuring the half-life of an isotope.
13.2.8 Solve problems involving radioactive half-life.

A nucleus that can decay into another is said to be unstable; the more energy released, the more unstable the nucleus is.

W To view a simulation of random decay, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 7.4.

It is not possible to say exactly when an unstable nucleus will decay, but if you compare two nuclei, then the one that is most unstable is most likely to decay first. It is like watching two leaves on a tree – you don't know which will fall first, but the brown crinkly one will probably fall before the green one. Staying with the tree analogy, it is also the case that the number of leaves that fall in an hour will be greater if there are more leaves on the tree. In the same way, the number of decays per unit time of an amount of unstable material is proportional to the number of nuclei.

$$\text{Rate of decay} \propto \text{number of nuclei.}$$

The exponential decay curve

If we start with 100 unstable nuclei, then as time progresses and the nuclei decay, the number of nuclei remaining decreases. As this happens, the rate of decay also decreases. If we plot a graph of the number of nuclei against time, the gradient (rate of decay) starts steep but gets less steep with time.

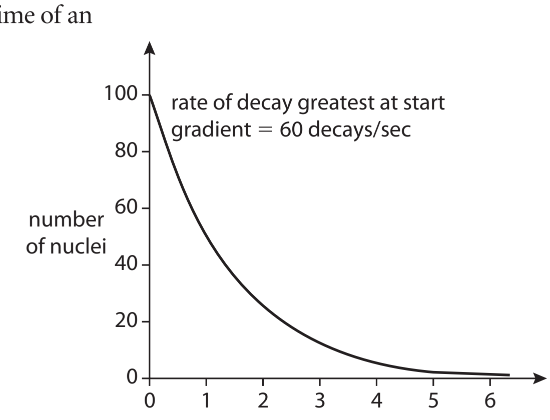
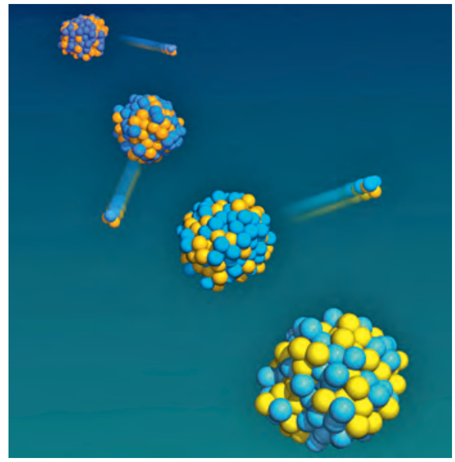


Figure 7.43 The shape of the decay curve is an exponential.

This curve tends towards zero but will never get there (it's an asymptote), so it is impossible to say at what time the last nucleus will decay. However, we can say how long it will take for *half* of the nuclei to decay – this is called the *half-life*.

An artist's impression of nuclei randomly emitting alpha particles.

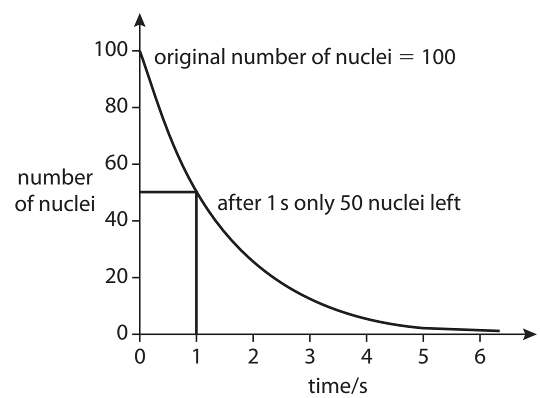


Half-life

The half-life is defined as the time taken for half of the nuclei in a sample to decay.

In Figure 7.44, you can see that the time taken for the original 100 nuclei to decay to 50 is 1 second. The half-life of this material is therefore 1second. You can also see that after each further second the number of nuclei keeps on halving.

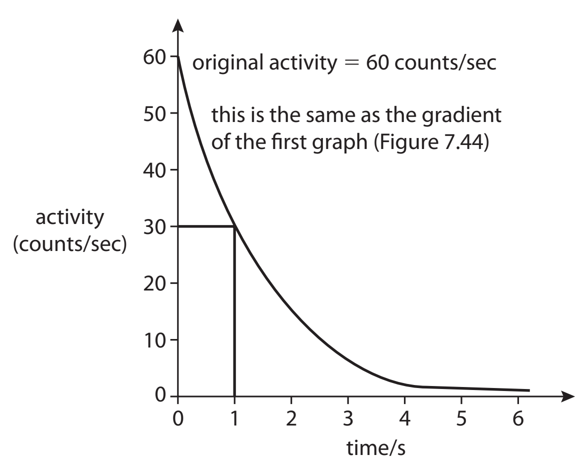
Figure 7.44 Decay curve showing half-life.



Activity

It is not very easy to count the number of undecayed nuclei in a sample – it is much easier to measure the radiation. Since the rate of decay is proportional to the number of nuclei, a graph of the rate of particle emission against time will have the same shape.

Figure 7.45 Activity *vs* time.



Worked examples

Cobalt-60 decays by beta emission and has a half-life of approximately 5 years. If a sample of cobalt-60 emits 40 beta particles per second, how many will the same sample be emitting in 15 years time?

Solution

After 5 years the activity will be 20 particles per second.

After another 5 years it will be 10 particles per second.

Finally after a further 5 years it will emit 5 particles per second (5 Bq).

Becquerel (Bq)
The becquerel is the unit of activity, measured as counts per second.

Carbon dating
There are two isotopes of carbon in the atmosphere: ^{14}C and ^{12}C. ^{12}C is stable but ^{14}C is radioactive with a half-life of 6000 years. However, it is made as quickly as it decays, so the ratio of ^{14}C to ^{12}C is always the same. As plants grow, they continuously absorb carbon from the atmosphere, so the ratio of ^{14}C to ^{12}C is also constant in them. When plants die, the ^{14}C decays, so by measuring the ratio of ^{14}C to ^{12}C, the age of a piece of dead organic matter can be found.

Exercises

24 ^{17}N decays into ^{17}O with a half-life of 4 s. How much ^{17}N will remain after 16 s, if you start with 200 g?

25 ^{11}Be decays into ^{11}B with a half-life of 14 s. If the ^{11}Be emits 100 particles per second, how many particles will it emit after 42 s?

26 A sample of dead wood contains $\frac{1}{16}$ of the amount of ^{14}C that it contained when alive. If the half-life of ^{14}C is 6000 years how old is the sample?

The exponential decay equation

If the time for a decay is a whole number of half-lives, it is easy to calculate how much of a sample has decayed. However, if not then you need to use the decay equation.

We have seen that radioactive decay is a random process, therefore: the rate of decay is proportional to the number of undecayed nuclei. We can write this a differential equation:

$$\frac{dN}{dt} = -\lambda N$$

where λ is the decay constant.

You have probably learnt how to solve equations like this in mathematics, using the method of 'separation of variables' and integrating between time 0 when $N = N_0$ and time t when $N = N_t$.

The solution is:

$$N_t = N_0 e^{-\lambda t}$$

This is called an *exponential decay equation*.

Use of spreadsheets
You can use a spreadsheet like Excel to plot the decay curve and see what happens when you change $\boldsymbol{\lambda}$. The exponential equation in Excel is exp().

The decay constant

The decay constant, λ tells us how quickly the material will decay. This is illustrated by the two lines in Figure 7.46.

The blue line has the bigger decay constant so decays more rapidly.

From the equation $\frac{dN}{dt} = -\lambda N$ we can see that the units of λ are s^{-1}.

If we only have one nucleus then $\frac{dN}{dt} = -\lambda$. If this is large then the rate of decay of that nucleus is high. But with one nucleus you can't really talk about the rate of decay – instead we really mean the *probability* of it decaying, so λ is the probability of a nucleus decaying in a second.

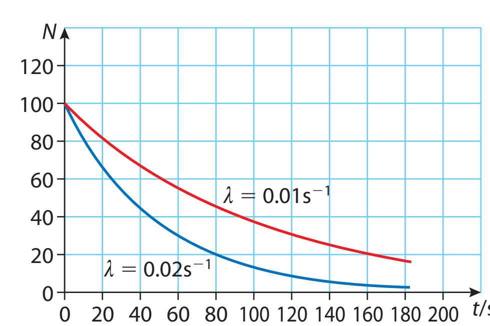


Figure 7.46 Two exponential decays with different decay constants.

Relationship between λ and half-life

The half-life, $t_{\frac{1}{2}}$ is the time taken for the number of nuclei to decay to half the original value. So if the original number of nuclei $= N_0$ then after $t_{\frac{1}{2}}$ seconds there will be $\frac{N_0}{2}$ nuclei. If we substitute these values into the exponential decay equation, we get:

$$\frac{N_0}{2} = N_0 e^{-\lambda t_{\frac{1}{2}}}$$

Cancelling N_0

$$\tfrac{1}{2} = e^{-\lambda t_{\frac{1}{2}}}$$

Taking natural logs of both sides gives

$$\ln\left(\tfrac{1}{2}\right) = -\lambda t_{\frac{1}{2}}$$

This is the same as

$$\ln 2 = \lambda t_{\frac{1}{2}}$$

So

$$t_{\frac{1}{2}} = \frac{0.693}{\lambda}$$

So if we know the decay constant, we can find the half-life or vice versa.

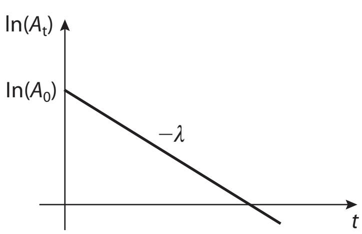


Figure 7.47 Graph used to determine the half-life.

Activity equation

The activity, A, of a sample is the number of particles (α, β or γ) emitted per unit time. This is the same as the number of decays per unit time $\left(\frac{dN}{dt}\right)$ so the equation for how the activity varies with time is also exponential.

$$A_t = A_0 e^{-\lambda t}$$

You can use 'decays per second' as the unit of activity, but the SI unit is the becquerel (Bq). This is the amount of a given material that emits 1 particle per second, so if you have a piece of material that has an activity of 5 Bq then it will emit 5 particles per second.

Measuring half-life

There are two ways to measure the half-life of an isotope; either measure the activity over a period of time or the change in number of nuclei.

In the school laboratory you might do an experiment using an isotope with a half-life of a few minutes, such as protactinium. In this case it is possible to measure the count rate every 10 seconds for a couple of minutes and find the half-life from that. The best way to analyse the data is to plot $\ln(A_t)$ against t. If you take logs of the activity equation you get:

$$\ln(A_t) = \ln(A_0) - \lambda t$$

So if you plot $\ln(A_t)$ against t you will get a straight line with $-\lambda$ as the gradient.

If the half-life is very long then you would have to wait a very long time before the count rate changed. In this case, a mass spectrometer could be used to find the number of different nuclei present at different times rather than the activity.

Radioactive dating

Potassium-argon dating

Potassium-argon dating is used to date rocks. Potassium-40 decays to argon-40 with a half-life of 1.26×10^9 years. When rock containing potassium is molten (hot liquid) any argon that is formed will bubble up to the surface and leave the rock, but when the rock solidifies the argon is trapped. If a rock sample is heated the argon atoms are released and can be counted with a mass spectrometer. If you also count the potassium nuclei you can calculate the age of the rock.

Radioactive dating brings science into conflict with some religions, as the age of Earth calculated from radioactive dating doesn't agree with the age given in religious teaching. How can these conflicts be resolved?

Carbon dating

Carbon dating is only used for organic material up to about 60 000 years old. There are two isotopes of carbon present in living material: carbon-12 and carbon-14. The percentage of $^{14}C/^{12}C$ is always the same in living things (about $10^{-10}\,\%$) but when they die, the ^{14}C begins to decay. As it decays, the percentage decreases exponentially. If you measure the percent of $^{14}C/^{12}C$ in a dead specimen, you can calculate its age using the formula

$$\%_{\text{now}} = \%_{\text{originally}} \times e^{-\lambda t}$$

where the original percentage is $10^{-10}\,\%$.

^{14}C half-life
The half-life of ^{14}C is 6000 years. It can't be used to date samples older than about 10 half-lives (60 000 years).

Worked example

Protactinium has a half-life of 70 s. A sample has an activity of 30 Bq. Calculate its activity after 10 minutes.

Solution

We are going to use the equation $A_t = A_0 e^{-\lambda t}$. First we need to find the decay constant $\lambda = \frac{0.693}{t_{\frac{1}{2}}} = 9.9 \times 10^{-3}\,\text{s}^{-1}$

The activity after 700 s is now $= 30 \times e^{-9.9 \times 10^{-3} \times 700} = 0.03$ Bq

Exercises

27 A sample has activity 40 Bq. If it has a half-life of 5 mins, what will its activity be after 12 mins?

28 The activity of a sample of strontium-90 decreases from 20 Bq to 15.7 Bq in 10 years. What is the half-life of strontium 90?

29 A sample of rock is found to contain 7×10^{10} potassium-40 nuclei and 3×10^{10} argon nuclei. Calculate

(a) the original number of potassium nuclei present in the sample

(b) the age of the sample.

30 A sample of dead wood contains $0.9 \times 10^{-10}\,\%$ ^{14}C compared to ^{12}C. What is the age of the sample?

31 Cobalt-60 has a half-life of 5.27 years. Calculate

(a) the half-life in s

(b) the decay constant in s^{-1}

(c) the number of atoms in one gram of ^{60}Co

(d) the activity of l gram of ^{60}Co

(e) the amount of ^{60}Co in a sample with an activity of 50 Bq.

Hint: You don't have to convert everything into seconds. If half-life is in years then $\boldsymbol{\lambda}$ is year^{-1}.

7.8 Nuclear reactions

Assessment statements

7.3.1 Describe and give an example of an artificial (induced) transmutation.
7.3.2 Construct and complete nuclear equations.
7.3.8 Describe the processes of nuclear fission and nuclear fusion.
7.3.9 Apply the graph in 7.3.6 to account for the energy release in the processes of fission and fusion.
7.3.10 State that nuclear fusion is the main source of the Sun's energy.
7.3.11 Solve problems involving fission and fusion reactions.

Transmutation

We have seen how nuclei can change from one form to another by emitting radioactive particles. It is also possible to change a nucleus by adding nucleons. These changes or transmutations can occur naturally, as in the production of nitrogen from carbon in the atmosphere, or can be artificially initiated by bombarding a target material with high-energy particles.

Transmutation of nitrogen into carbon

Nuclide/particle	Mass (u)
^{14}N	14.0031
neutron	1.008664
^{14}C	14.003241
proton	1.007825

$$^{14}_{7}N + ^{1}_{0}n \rightarrow ^{14}_{6}C + ^{1}_{1}p$$

In this reaction, a nitrogen nucleus absorbs a neutron and gives out a proton.

Note: In all nuclear reactions, the nucleon number and proton number must balance.

Comparing the masses before and after the reaction:

Initial mass − final mass = 0.000698 u

This loss of mass must have been converted to the KE of the proton.

$E = 931.5 \times 0.000698 = 0.65$ MeV

Remember: 1 u = 931.5 MeV.

Exercises

32 In the following transmutations, fill in the missing nucleon and proton numbers.

(a) $^{4}_{2}He + ^{16}_{8}O \rightarrow ^{?}_{9}F + ^{1}_{?}H + ^{1}_{0}n$

(b) $^{4}_{2}He + ^{121}_{51}Sb \rightarrow ^{?}_{?}I + 2^{1}_{0}n$

(c) $^{2}_{1}H + ^{17}_{?}N \rightarrow ^{?}_{8}O + ^{1}_{0}n$

(d) $^{31}_{15}P \rightarrow ^{?}_{15}P + ^{1}_{0}n + \gamma$

33 (a) Using the table, calculate the change in mass in Question 32 (a).

Nuclide/particle	Mass (u)
^{4}He	4.002 603
^{16}O	15.994 914
^{18}F	18.998 403
^{1}H	1.007 825
neutron	1.008 664

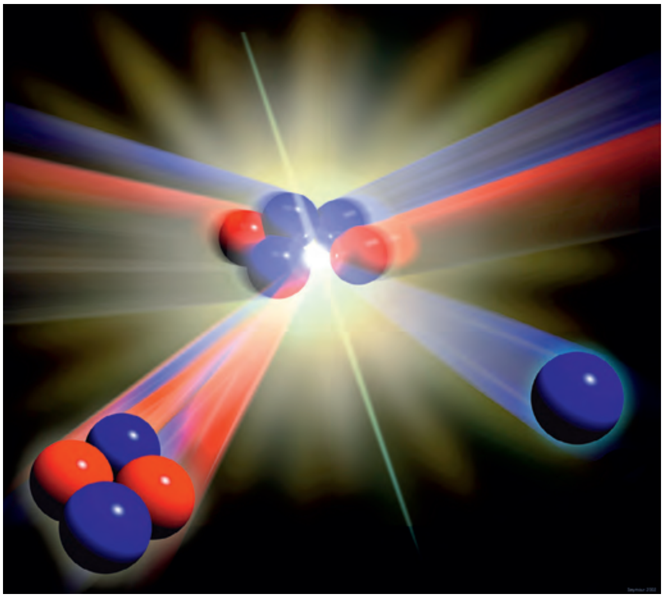
You will notice that this is a negative number, which means that energy is required to make it happen. This energy is supplied by a particle accelerator used to accelerate the helium nucleus.

(b) From your previous answer calculate the KE of the He nucleus.

Nuclear fusion

Nuclear fusion is the joining up of two small nuclei to form one big one.

If we look at the BE/nucleon *vs* nucleon number curve (Figure 7.48), we see that the line initially rises steeply. If you were to add two $^{2}_{1}H$ nuclei to get one $^{4}_{2}He$ nucleus, then the He nucleus would have more BE per nucleon.



An artist's impression of the fusion of ^{2}H and ^{3}H to form ^{4}He.

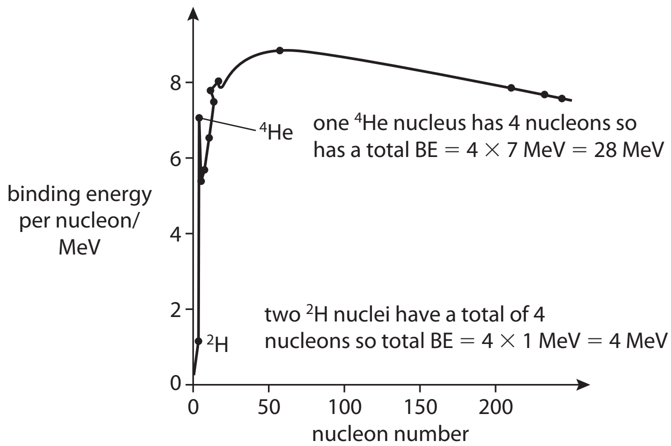


Figure 7.48 BE/nucleon *vs* nucleon number curve, showing fusion possibility.

If we add up the total BE for the helium nucleus, it has 24 MeV more BE than the two hydrogen nuclei. This means that 24 MeV would have to be released; this could be by the emission of gamma radiation.

Nuclide	Mass in u
^{1}H	1.007825
^{2}H	2.014 101
^{3}H	3.016 049
^{3}He	3.016 029
^{4}He	4.002 603
^{o}n	1.008 664

Worked example

Calculate the energy released by the following reaction:

$$^{2}_{1}H + ^{3}_{1}H \rightarrow ^{4}_{2}He + ^{1}_{0}n$$

Solution

If the masses are added, we find that the mass of the original nuclei is greater than the mass of the final ones. This mass has been converted to energy.

Mass difference $= 0.018883$ u

1u is equivalent to 931.5 MeV so energy released $= 17.6$ MeV

It wouldn't be possible to conserve both KE and momentum if two fast moving nuclei collided and fused together. That's why all the reactions result in two particles not one big one.

Each small nucleus has a positive charge so they will repel each other. To make the nuclei come close enough for the strong force to pull them together, they must be thrown together with very high velocity. For this to take place, the matter must either be heated to temperatures as high as the core of the Sun (about 13 million kelvin) or the particles must be thrown together in a particle accelerator.

The fusion reaction produces a lot of energy per unit mass of fuel and much research has been carried out to build a fusion reactor. You can read more about this in Chapter 8.

Exercises

34 Use the data in the table above to calculate the change in mass and hence the energy released in the following examples of fusion reactions:

(a) $^{2}_{1}H + ^{2}_{1}H \rightarrow ^{3}_{2}He + ^{1}_{0}n$

(b) $^{2}_{1}H + ^{2}_{1}H \rightarrow ^{3}_{1}H + ^{1}_{1}p$

(c) $^{2}_{1}H + ^{3}_{2}He \rightarrow ^{4}_{2}He + ^{1}_{1}p$

Nuclear fission

Looking at the right-hand side of the graph we see that if one large nucleus is split into two smaller ones, then the total BE would again be increased. This reaction forms the basis of the nuclear reactor that you will learn more about in Chapter 8.

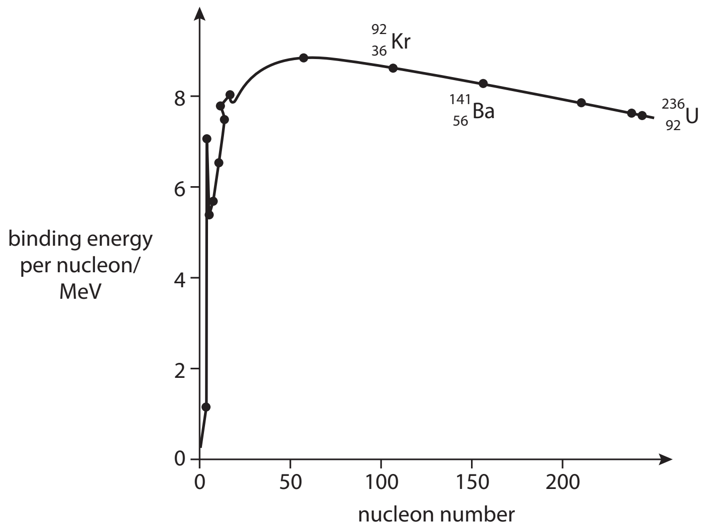


Figure 7.49 BE/nucleon *vs* nucleon number curve showing fission possibility.

Worked example

Find the energy released if uranium-236 splits into krypton-92 and barium-141.

Solution

BE of ^{236}U $= 236 \times 7.6$

$= 1793.6$ MeV

BE of ^{141}Ba $= 141 \times 8$

$= 1128$ MeV

BE of ^{92}Kr $= 92 \times 8.2$

$= 754.4$ MeV

Gain in BE $= (1128 + 754.4) - 1793.6$

$= 88.8$ MeV

Since this leads to a release of energy, this process is possible.

● **Examiner's hint:** Use the graph to find the BE per nucleon.

Exercises

Table of nuclear masses			
Z	Symbol	*A*	Mass (u)
92	U	233	233.039628
92	U	236	236.045563
42	Mo	100	99.907476
50	Sn	126	125.907653
56	Ba	138	137.905233
36	Kr	86	85.910615
0	n	1	1.008664

Use the table above to answer the following questions.

35 If ^{236}U splits into ^{100}Mo and ^{126}Sn, how many neutrons will be produced? Calculate the energy released in this reaction.

36 ^{233}U splits into ^{138}Ba and ^{86}Kr plus 9 neutrons. Calculate the energy released when this takes place.

● **Examiner's hint:** To answer these questions you must find the difference in mass between the original nucleus and the products. To convert to MeV, simply multiply by 931.5.

Practice questions

1 This question is about nuclear reactions.

(a) Complete the table below, by placing a tick (✓) in the relevant columns, to show how an increase in each of the following properties affects the rate of decay of a sample of radioactive material.

Property	Effect on rate of decay		
	increase	decrease	stays the same
temperature of sample			
pressure on sample			
amount of sample			

(2)

Radium-226 ($^{226}_{88}$Ra) undergoes natural radioactive decay to disintegrate spontaneously with the emission of an alpha particle (α-particle) to form radon (Rn). The masses of the particles involved in the reaction are

radium: 226.0254 u
radon: 222.0176 u
α-particle: 4.0026 u

(b) **(i)** Complete the nuclear reaction equation below for this reaction.

$$^{226}_{88}\text{Ra} \rightarrow \, ^{\cdots\cdots}_{\cdots\cdots}\cdots\cdots + \, ^{\cdots\cdots}_{\cdots\cdots}\text{Rn}$$ (2)

(ii) Calculate the energy released in the reaction. (3)

(c) The radium nucleus was stationary before the reaction.

(i) Explain, in terms of the momentum of the particles, why the radon nucleus and the α-particle move off in opposite directions after the reaction. (3)

(ii) The speed of the radon nucleus after the reaction is v_R and that of the α-particle is v_α. Show that the ratio $\frac{v_\alpha}{v_R}$ is equal to 55.5. (3)

(iii) Using the ratio given in (ii) above, deduce that the kinetic energy of the radon nucleus is much less than the kinetic energy of the α-particle. (3)

(d) Not all of the energy of the reaction is released as kinetic energy of the α-particle and of the radon nucleus. Suggest **one** other form in which the energy is released. (1)

Another type of nuclear reaction is a fusion reaction. This reaction is the main source of the Sun's radiant energy.

(e) **(i)** State what is meant by a *fusion reaction*.

(ii) Explain why the temperature and pressure of the gases in the Sun's core must both be very high for it to produce its radiant energy. (5)

(Total 25 marks)

2 This question is about nuclear reactions.

(a) **(i)** Distinguish between *fission* and *radioactive decay*. (4)

A nucleus of uranium-235 ($^{235}_{92}$U) may absorb a neutron and then undergo fission to produce nuclei of strontium-90 ($^{90}_{38}$Sr) and xenon-142 ($^{142}_{54}$Xe) and some neutrons. The strontium-90 and the xenon-142 nuclei both undergo radioactive decay with the emission of β^- particles.

(ii) Write down the nuclear equation for this fission reaction. (2)

(iii) State the effect, if any, on the mass number (nucleon number) and on the atomic number (proton number) of a nucleus when the nucleus undergoes β^- decay.

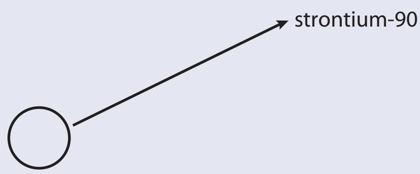
Mass number: Atomic number: (2)

The uranium-235 nucleus is stationary at the time that the fission reaction occurs. In this fission reaction, 198 MeV of energy is released. Of this total energy, 102 MeV and 65 MeV are the kinetic energies of the strontium-90 and xenon-142 nuclei respectively.

(b) **(i)** Calculate the magnitude of the momentum of the strontium-90 nucleus. (4)

(ii) Explain why the magnitude of the momentum of the strontium-90 nucleus is not exactly equal in magnitude to that of the xenon-142 nucleus. (2)

On the diagram right, the circle represents the position of a uranium-235 nucleus before fission. The momentum of the strontium-90 nucleus after fission is represented by the arrow.



(iii) On the diagram above, draw an arrow to represent the momentum of the xenon-142 nucleus after the fission. (2)

(c) In a fission reactor for the generation of electrical energy, 25% of the total energy released in a fission reaction is converted into electrical energy.

(i) Using the data in **(b)**, calculate the electrical energy, in joules, produced as a result of nuclear fission of one nucleus. (2)

(ii) The specific heat capacity of water is 4.2×10^3 J kg^{-1}. Calculate the energy required to raise the temperature of 250 g of water from 20 °C to its boiling point (100°C). (3)

(iii) Using your answer to **(c)(i)**, determine the mass of uranium-235 that must be fissioned in order to supply the amount of energy calculated in **(c)(ii)**. The mass of a uranium-235 atom is 3.9×10^{-25} kg. (4)

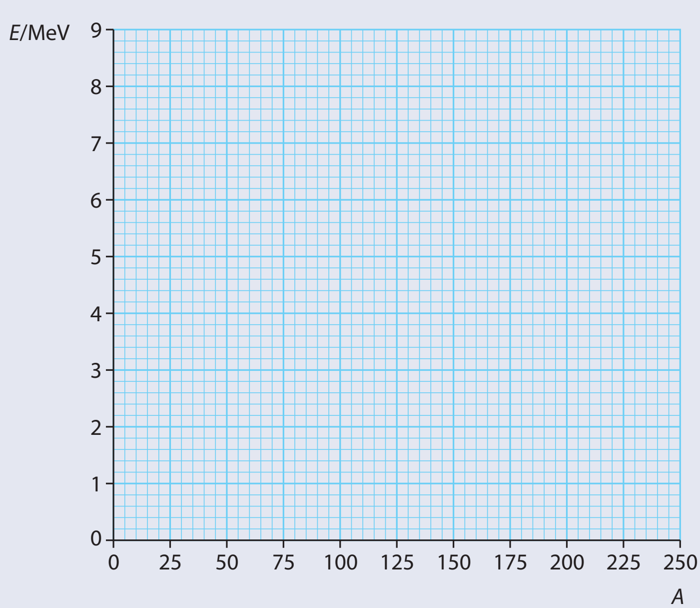
(Total 25 marks)

3 This question is about nuclear binding energy.

(a) **(i)** Define *nucleon*. (1)

(ii) Define *nuclear binding energy of a nucleus*. (1)

The axes below show values of nucleon number *A* (horizontal axis) and average binding energy per nucleon *E* (vertical axis). (Binding energy is taken to be a positive quantity.)



(b) Mark on the E axis above, the approximate position of

(i) the isotope ${}^{56}_{26}\mathrm{Fe}$ (label this F). (1)

(ii) the isotope ${}^{2}_{1}\mathrm{H}$ (label this H). (1)

(iii) the isotope ${}^{238}_{92}\mathrm{U}$ (label this U). (1)

(c) Using the grid in part **(a)**, draw a graph to show the variation with nucleon number A of the average binding energy per nucleon E. (2)

(d) Use the following data to deduce that the binding energy per nucleon of the isotope ${}^{3}_{2}\mathrm{He}$ is 2.2 MeV.

nuclear mass of ${}^{3}_{2}\mathrm{He}$ = 3.01603 u

mass of proton = 1.00728 u

mass of neutron = 1.00867 u (3)

In the nuclear reaction

$${}^{2}_{1}\mathrm{H} + {}^{2}_{1}\mathrm{H} \rightarrow {}^{3}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n}$$

energy is released.

(e) **(i)** State the name of this type of reaction. (1)

(ii) Use your graph in **(c)** to explain why energy is released in this reaction. (2)

(Total 13 marks)

4 This question is about the wave nature of matter.

(a) Describe the concept of matter waves and state the de Broglie hypothesis (3)

(b) An electron is accelerated from rest through a potential difference of 850 V. For this electron

(i) calculate the gain in kinetic energy. (1)

(ii) deduce that the final momentum is 1.6×10^{-23} Ns. (2)

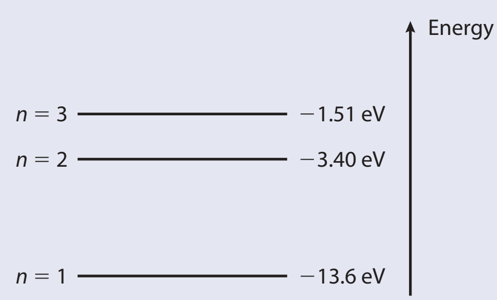
(iii) determine the associated de Broglie wavelength.

(Electron charge $e = 1.6 \times 10^{-19}$ C, Planck constant $h = 6.6 \times 10^{-34}$ J s) (2)

(Total 8 marks)

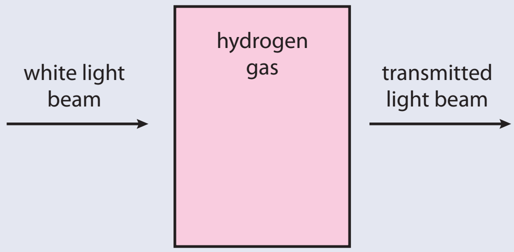
5 This question is about the Bohr model of the hydrogen atom.

(a) The diagram below shows the three lowest energy levels of a hydrogen atom as predicted by the Bohr model.



State **two** physical processes by which an electron in the ground state energy level can move to a higher energy level state. (2)

(b) A parallel beam of white light is directed through monatomic hydrogen gas as shown in the diagram opposite. The transmitted light is analysed.



White light consists of photons that range in wavelength from approximately 400 nm for violet to 700 nm for red light.

(i) Determine that the energy of photons of light of wavelength 658 nm is about 1.89 eV. (2)

(ii) The intensity of light of wavelength 658 nm in the direction of the transmitted beam is greatly reduced. Using the energy level diagram in **(a)** explain this observation. (3)

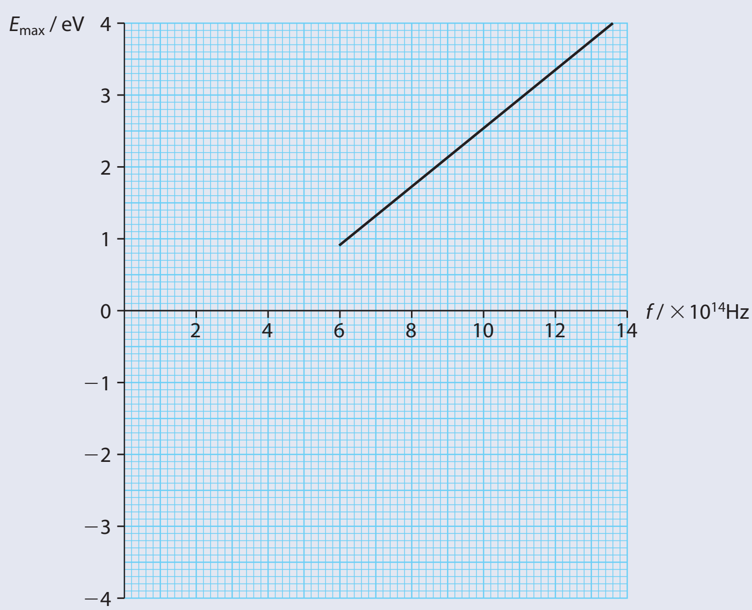
(iii) State **two** ways in which the Schrödinger model of the hydrogen atom differs from that of the Bohr model. (2)

(Total 9 marks)

6 This question is about the photoelectric effect.

(a) State **one** aspect of the photoelectric effect that **cannot** be explained by the wave model of light. Describe how the photon model provides an explanation for this aspect. (2)

Light is incident on a metal surface in a vacuum. The graph below shows the variation of the maximum kinetic energy E_{max} of the electrons emitted from the surface with the frequency f of the incident light.



(b) Use data from the graph to determine

(i) the threshold frequency. (2)

(ii) a value of the Planck constant. (2)

(iii) the work function of the surface. (2)

The threshold frequency of a different surface is 8.0×10^{14} Hz.

(c) On the axes opposite, draw a line to show the variation with frequency f of the maximum kinetic energy E_{max} of the electrons emitted. (2)

(Total 10 marks)

7 This question is about radioactive decay and the age of rocks.

A nucleus of the radioactive isotope potassium-40 decays into a stable nucleus of argon-40.

(a) Complete the equation below for the decay of a potassium-40 nucleus.

$$^{40}_{19}\text{K} \longrightarrow {}^{40}_{18}\text{Ar} +$$ (2)

A certain sample of rocks contains 1.2×10^{-6} g of potassium-40 and 7.0×10^{-6} g of trapped argon-40 gas.

(b) Assuming that all the argon originated from the decay of potassium-40 and that none has escaped from the rocks, calculate what mass of potassium was present when the rocks were first formed. (1)

The half-life of potassium-40 is 1.3×10^9 years.

(c) Determine

(i) the decay constant of potassium-40. (2)

(ii) the age of the rocks. (2)

(Total 7 marks)

8 This question is about charged particles in a magnetic field.

A beam of singly ionized atoms moving at speed v enters a region of magnetic field strength B as shown below.

The magnetic field is directed into the plane of the paper. The ions follow a circular path,

(a) Deduce that the radius r of the circular path is given by

$$r = \frac{mv}{Bq}$$

where m and q are the mass and charge respectively of the ions. (2)

In one particular experiment, the beam contains singly ionized neon atoms all moving at the same speed. On entering the magnetic field, the beam divides in two. The path of the ions of mass 20 u has radius 15.0 cm.

(b) Calculate, in terms of u, the mass of the ions having a path of radius 16.5 cm. (2)

The atomic number (proton number) of neon is 10.

(c) State the number of protons and neutrons in each type of neon ion. (2)

(Total 6 marks)

8 Energy, power and climate change

8.1 Energy degradation and power generation

Assessment statements

8.1.1 State that thermal energy may be completely converted to work in a single process, but that continuous conversion of this energy into work requires a cyclical process and the transfer of some energy from the system.

8.1.2 Explain what is meant by degraded energy.

8.1.3 Construct and analyse energy flow diagrams (Sankey diagrams) and identify where the energy is degraded.

8.1.4 Outline the principal mechanisms involved in the production of electrical power.

In this section we will look at how we use energy. You already know from the law of conservation of energy that energy cannot be created or destroyed. So we cannot make energy, we can only change it from one form to another. There are only two ways to transfer energy from one body to another – either by doing work or by transfering thermal energy. Before looking at the different ways that we utilize sources of energy, we will consider the basic physical principles behind converting the energy possessed by different sources into useful work.

Fuels

When we think of sources of energy we often think of fuels such as coal, oil and nuclear fuel. A fuel is a substance that can release energy by changing its chemical or nuclear structure. For example, when a piece of coal burns it is changing its chemical structure. As far as we need to know, that means that the atoms are changing their positions; this leads to a reduction in their PE and hence an increase in their KE. The average KE is related to the temperature of the material so the coal gets hot. This is obviously great if we need heat but if we want to use the energy to turn the wheels of a machine we need an engine – coal can't do work on its own. There are many different types of engine but they all work on the same physical principles. To demonstrate this we will consider a very simple but rather impractical engine that uses a hot air balloon.

The hot air balloon engine

Figure 8.1 shows the hot air balloon engine; it might never have been made but it would work. The hot air in the balloon is less dense than the surrounding cold air so the balloon floats up in the same way that a cork will float to the top of a bowl of water. As the balloon rises, a rope attached to it turns the reel, producing

Figure 8.1 Underneath the balloon is a flame which heats the air in the balloon and this drives the engine.

mechanical energy, so the balloon is doing work on the reel. The heat produced by the burning fuel is being converted to energy that could be used to turn a machine or drive a car.

The problem is that once the rope has all been used up the machine will stop. To make a continuous process you have to get the balloon down and to do that it has to lose thermal energy. If the balloon loses thermal energy and cools down it will fall back to the ground and the process can be repeated. This is true for all engines; they can only work if some thermal energy is transferred to the surroundings. You may think that if you could catch this energy then you could use it again and create a machine that kept on going forever without requiring any input of energy, a perpetual motion machine, but this is not possible.

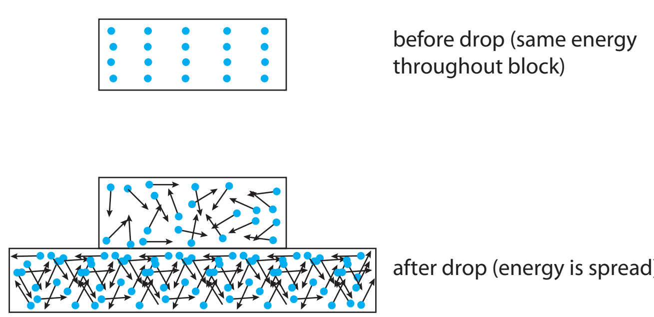
The second law of thermodynamics

For many years scientists tried to make engines that would turn thermal energy into useful work without loss of heat. Eventually they decided that their failure was not due to bad experimental method but that it simply wasn't possible. A law was written to take this into account.

It is impossible to take heat from a hot object and use it all to do work without losing some to the surroundings.

The reason for this is all to do with the nature of matter and the way it behaves in our universe. We have seen that matter is made up of molecules and if you transfer thermal energy to a body then the molecules start to move faster. If we take the example of a block (whose temperature is 0 K) falling to the ground as in Figure 8.2, then we start with the PE of the block held above the ground and end up with thermal energy when it hits the ground. If we compare the molecules before and after, then we see that before, all the molecules had about the same energy due to their position above the ground, but afterwards the energy is spread out between the molecules in the ground and the block. Whenever energy is transferred it always spreads out, and once it spreads out you can't get it back. This is why engines cannot convert heat completely to do work and also why it is not possible to make a perpetual motion machine: energy is always lost to the surroundings.

Figure 8.2 As time progresses the energy from the dropped block spreads further and further through the ground.



Degradation of energy

We can see from the example above that energy becomes more spread out or disordered. This is called the *degradation of energy*. Thermal energy is the most degraded form of energy and we have seen that to convert thermal energy into mechanical energy we must always transfer some of the thermal energy to the cold surroundings. Once energy becomes thermal energy we can never get it all back.

All energy eventually turns to heat. Taking this to extremes means that the end of the universe will be when everything is at the same temperature; it will then be impossible to do any work.

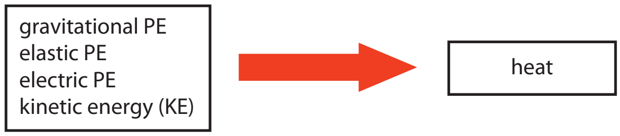


Figure 8.3 Energy is classified in terms of how degraded it is. As time goes on all energy turns into heat.

One minute we were talking about engines and the next about the end of the universe. This is an example of how physical laws have many applications.

Examples of energy transfer

There are many examples of devices that transfer energy from one form to another. According to the second law, energy transfer must always lead to some thermal energy being transferred to the surroundings. Let us test this with a few examples:

Light bulb – Converts electrical energy to light + heat.

Electric motor – Converts electrical energy to mechanical + heat.

Battery – Converts chemical energy to electrical + heat

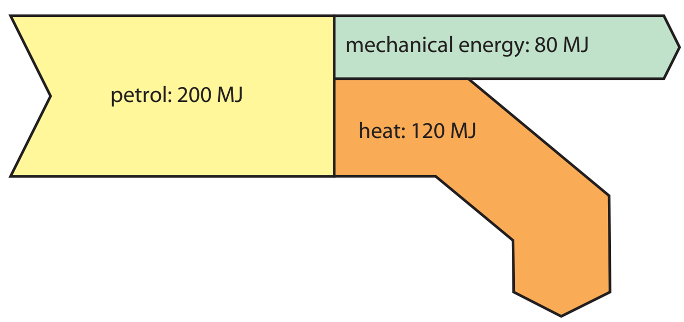
Car engine – Converts chemical energy to mechanical + heat

Solar cell – Converts light to electrical + heat

Sankey diagrams

Sankey diagrams are used to visualize the flow of something, for instance the flow of water down a river or the flow of energy as it changes from one form to another. In the rest of this chapter Sankey diagrams will be used to represent different ways of producing useful energy. Here are some examples:

Figure 8.4 Sankey diagram for a car engine.



The car engine

A car engine converts the chemical energy in petrol to the KE of the car plus heat. The Sankey diagram shows 200 MJ of chemical energy in the petrol coming into the car and changing to 80 MJ of mechanical energy and 120 MJ of heat.

Electric motor

An electric motor uses the energy flow from a battery to drive the motor.

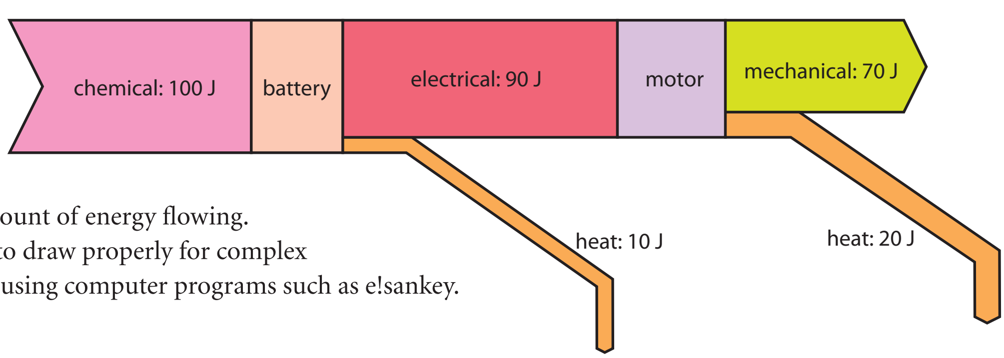


Figure 8.5 Sankey diagram for an electric motor.

In Sankey diagrams the widths of the elements are proportional to the amount of energy flowing. This can be quite difficult to draw properly for complex systems but is made easier using computer programs such as e!sankey.

Exercises

1 Draw a Sankey diagram for an electric light bulb.

2 Draw a Sankey diagram for a bicycle dynamo producing the electricity to illuminate a light from mechanical energy.

Generating electricity

Electricity is a very useful form of energy because it can be sent from one place to another through wires. Taking the example of heating a house, it is possible to heat a house using coal; however coal must first be transported from the coal mine to the house. This uses up a lot of energy, since trucks must be driven to each house every time they need more. It is much better to use the coal to produce electricity and send that electricity to the house through wires. Electricity can be produced from many different types of energy, and all of these methods make use of a generator to convert mechanical energy to electrical energy.

Electrical generator

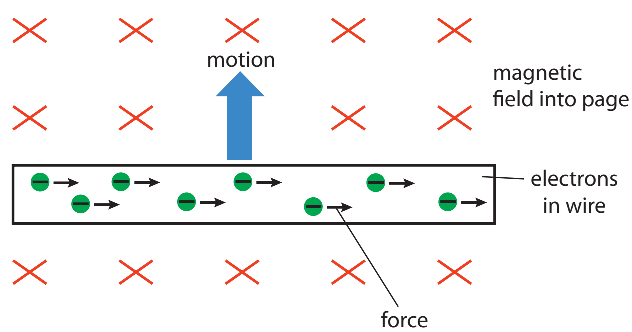


Figure 8.6 As electrons move upwards with the wire they will experience a force to the right (use Fleming's left hand rule but remember the electrons move opposite to current flow so current is downwards).

The coils of an electricity generator. Magnets fitted to a cylinder rotate inside the coils causing electricity to be induced.

In Chapter 6 you learnt about the connection between electricity and magnetism: if a charged object moves in a magnetic field it experiences a *force*. Wires contain charged objects, electrons, so if a wire moves in a magnetic field, the electrons inside it experience a force causing them to move to one end of the wire. This causes a potential difference along the wire which can be used to create current in a circuit. We say that current has been induced in the wire.

A generator uses the same principle, but instead of the wire moving in a straight line, a coil rotates in the field. As the coil rotates, its sides are sometimes moving up and sometimes down through the field. This results in a current that keeps changing direction, called an *alternating current* or AC for short. To prevent the wires from getting twisted when the coil rotates, there must be sliding contacts between the coil and circuit. However, friction between these sliding components will eventually wear away the contacts. An alternative arrangement is for the magnets to rotate inside the coils, this also results in an alternating current but has no sliding connections.

The way that the current changes direction is difficult to imagine, but to see an animation, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 8.1.

This makes it very clear.

8.2 World energy sources

Assessment statements

8.2.1 Identify different world energy sources.
8.2.2 Outline and distinguish between renewable and non-renewable energy sources.
8.2.3 Define the *energy density* of a fuel.
8.2.4 Discuss how choice of fuel is influenced by its energy density.
8.2.5 State the relative proportions of world use of the different energy sources that are available.
8.2.6 Discuss the relative advantages and disadvantages of various energy sources.

Sources of energy

In this section we will look at the different sources of energy available on Earth that can be turned into useful work. From the principles of physics that we have just been looking at, we know that energy transfer always leads to the degradation of energy, so, if we want to do useful work, the energy source must be in a form that can be degraded or spread out. If we take the example of a room where everything is at the same temperature, then there is no way that we can make the energy more spread out than it is. However, if there is a red hot piece of coal in the middle of the room we can use the flow of energy from the coal to the room to power an engine. As this happens, the energy will become spread out in the room. So, sources of energy are either things that are hot like a volcano or the Sun, or things that contain ordered energy like the PE of the water in a reservoir or the chemical energy of coal.

Fuels

As mentioned before, a fuel is a substance that can release energy by changing its chemical or nuclear structure. There are several different sources of fuel available on Earth.

Energy density
The value of different fuels can be measured by their energy density. This is the amount of energy that can be extracted per kg of fuel. A fuel with high energy density costs less to transport than one with a low energy density. A comparison of different energy densities is given in Table 1 on page 271.

Coal

Coal is composed of plants that died millions of years ago. They were then covered by layers of sand and mud that squashed them into the hard black substance we call coal. For that reason it is called a fossil fuel. Whilst living, plants are able to convert the Sun's energy into plant matter by a process known as *photosynthesis*. In physics terms, the plant is reorganizing the atoms from the air plus some from the ground to make the large molecules that the plant is made of. This process requires energy and that energy comes from the Sun. To release the energy in coal, we have to burn it. This is problematic since it leads to the emission of carbon dioxide, one of the contributors to global warming.

Another problem with coal is that it has to be dug up from the ground. This was traditionally done by digging a deep hole until the coal was reached and then taking it out through the same hole (mining). With today's large machinery it is often cheaper to uncover the coal by removing all the rock above it (open cast mining), but this has a large environmental impact.

There is only a certain amount of coal in the ground and since it takes a long time to turn a tree into coal, the rate at which it is being used is much greater than the rate of production. Coal is, therefore, classified as a non-renewable source of energy. One day it will all be used up. However there is still a lot of coal left in the ground. In 2006 it was estimated that there was about 1×10^{15} kg of coal left, enough to meet world needs for 155 years.

Oil and gas

Although it was coal that fuelled the industrial revolution, since the 1950s oil has overtaken coal as the most important source of energy. Oil is another fossil fuel, formed from microscopic organisms that sank to the bottom of the sea when they died. Over a long period of time this organic matter was covered in sand that turned to rock and it was pressed into oil and gas. Oil and gas are easier to extract than coal since they are fluid and can be pumped up from where they lie underground. Sometimes extraction is not simple since a lot of oil reserves are in rock that is at the bottom of the sea. However, as technology has advanced, the methods of extracting oil have become more and more ingenious and it is now possible to drill for many kilometres and even go sideways from a platform that is floating in water many kilometres deep.

Like coal, oil is non-renewable and has to be burned to turn its stored energy into heat that can be used to power an engine. This makes oil another source of greenhouse gases.

Barrels
Oil is usually measured in barrels. One barrel is equivalent to 159 litres.

It is not possible to say how much oil there is left in the Earth because technology does not yet enable accurate predictions of how much remains in the known deposits. The information is also clouded by false claims. However, based only on statistics that were 90% sure, the estimate in 2003 was around 1×10^{14} litres.

Growing plants to meet our energy needs sounds like the perfect solution, but what would happen if all the agricultural areas in the world were used to produce fuel instead of food?

Wood and biomass

The reserves of fossil fuels are not renewable so one day will run out. Living plants convert sunlight plus carbon dioxide into plant matter, so can also be used as fuel. Wood is the most obvious example of this and has been used for heating and cooking since the time of primitive man. Sugarcane and many plants that produce oils can also be used to produce biofuel that can be used to run cars. It is possible to run a diesel car on the fat that you use to fry food in the kitchen. Food waste can also be used to produce fuel as can cow manure; however these are not high grade fuels.

These fuels have to be burned so that their energy can be used, which means that they also contribute to global warming. However, whilst the plants were growing, they were absorbing carbon dioxide and reducing the overall greenhouse effect.

Nuclear fuel

Unlike the other fuels mentioned above, the energy in nuclear fuel does not come from the Sun. Nuclear fuels are materials that have unstable nuclei, and when these nuclei split they give out energy. This was explained in Chapter 7. Nuclear fuel is a non-renewable resource and although it does not have to be burned and therefore does not contribute to the greenhouse effect, it gives rise to waste materials that are highly radioactive and therefore difficult to dispose of. Like coal, nuclear fuels such as uranium have to be extracted from the ground and then purified before use.

Fusion fuel

Although a test power plant has been successfully put into operation, *nuclear fusion* is not used for the commercial production of energy. The fuel is *hydrogen* that can be extracted from water. Although not renewable, there is a lot of water on the planet and only a little hydrogen is needed to produce a lot of energy (as can be seen from its energy density in Table 1).

Fuel	Energy density ($MJ\,kg^{-1}$)
Fusion fuel	300 000 000
Uranium 235	90 000 000
Petrol	46.9
Diesel	45.8
Biodiesel	42.2
Crude oil	41.9
Coal	32.5
Sugar	17.0
Wood	17.0
Cow dung	15.5
Household waste	10

Table 1 The energy density of materials shows the most suitable for use as fuels.

Other sources of energy derived from the Sun

Several of the fuels mentioned convert the energy radiated from the Sun into a material that can be burned. However, there are other ways that we can utilize the Sun's energy that are all renewable and produce no greenhouse gases.

Solar energy

The most obvious way of using the Sun's energy is to collect it directly. This can be done using mirrors to focus the radiation into a point. Once we have created a hot spot, the energy can be used to power an engine that can do work. Alternatively, the Sun's radiation can be turned directly into electricity using a photovoltaic cell.

Hydroelectric power

The Sun heats water in the sea and turns it into water vapour; this vapour produces clouds which lead to rain. When water rains onto a mountain the water has a high PE. This energy can be made to do work as the water runs down the mountain; this is the principle of the hydroelectric power station.

Wind power

When the Sun heats the atmosphere it causes the air to become less dense and move upwards. These air movements combined with the rotation of the Earth cause winds. The KE in moving air can be used to turn a turbine and produce electricity.

Wave power

The movement of the air can become very strong leading to storms; if these occur over the sea, then the water will be disturbed, leading to the production of waves. As waves spread out, so their energy spreads out, arriving at beaches all around the world. This energy can be used to power turbines and produce electricity. With present-day technology it is now becoming viable to use wave power for some isolated communities in areas of rough sea.

Sources of energy not derived from the Sun

Tidal power

Although the Sun has some influence on the tides they are mainly caused by the Moon. The reason for the tides is that the surface and bottom of the oceans are different distances from the Moon, so experience a different gravitational force, this causes the ocean to bulge outward on parts of the Earth that are closest and furthest away from the Moon. As the Earth rotates, these bulges travel round the Earth, causing the tides. This movement of water can be used to drive turbines and produce electricity.

Geothermal energy

In certain areas of the world parts of the Earth's crust move against each other, the friction between the rocks creates heat. When water runs into cracks in these rocks, it turns to steam and forms geysers on the Earth's surface. This explosive energy can be used to produce electricity.

Old Faithful, a geyser in the Yellowstone National Park, USA gives water some PE.

Worldwide consumption of energy

The following graphs show what percentages of the different energy sources are used worldwide. They take into account all use of fuel, including transport. The amount of energy used depends on how much time you measure the consumption for. In this case, it is per second, so the units used are watts (joules per second).

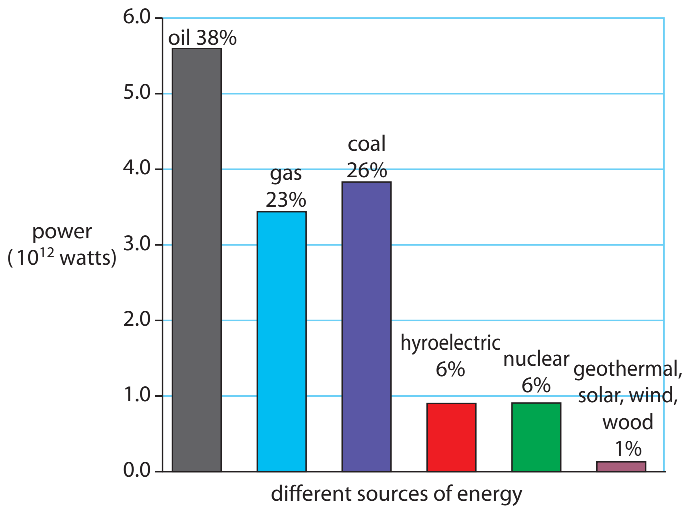


Figure 8.7 Worldwide usage of different sources of energy.

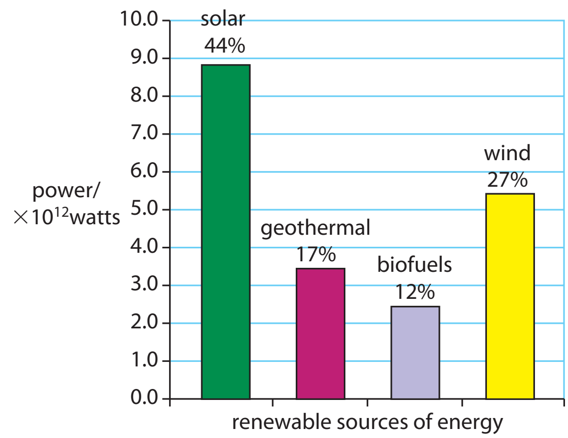


Figure 8.8 Worldwide usage of renewable sources of energy.

Worldwide consumption of electrical energy

Figure 8.9 shows the percentage of different fuels that are used to produce electricity alone.

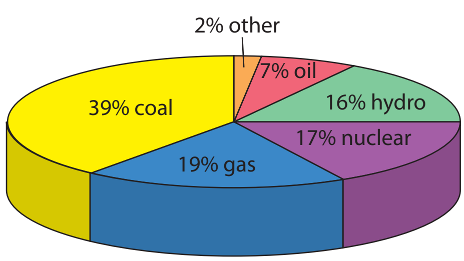


Figure 8.9 The percentage of different fuels used to produce electricity.

8.3 Fossil fuel power production

Assessment statements

8.3.1 Outline the historical and geographical reasons for the widespread use of fossil fuels.

8.3.2 Discuss the energy density of fossil fuels with respect to the demands of power stations.

8.3.3 Discuss the relative advantages and disadvantages associated with the transportation and storage of fossil fuels.

8.3.4 State the overall efficiency of power stations fuelled by different fossil fuels.

8.3.5 Describe the environmental problems associated with the recovery of fossil fuels and their use in power stations.

History of coal and oil from a physics perspective

It is known that coal has been used since the bronze age some 4000 years ago. In those days it was possible to find lumps of coal on the ground; however, early man did not have the same energy needs that we have today, so for them wood was a much more convenient source of fuel since it could be found almost anywhere. As technology advanced it became possible to extract coal from the ground by mining. However, this was a difficult and dangerous job and even though coal produced more heat than wood, it didn't become a major energy source until the industrial revolution.

In many developing countries wood is still the main source of fuel used in heating and cooking.

The invention of the steam engine

Although drawings of an early steam engine date back to the first century, they did not make a big impact until 1769 when James Watt invented a steam engine that could turn a wheel. Up to that point, steam engines had been able to move a beam up and down but the pistons did not move smoothly enough to turn a wheel. Suddenly, the possibility for powered transport became a reality and the first steam trains were on the tracks by the early 1800s. It is, of course, possible to use wood to power a steam engine but the energy density of wood is only half the energy density of coal, which means that twice as much wood is needed to do the same amount of work. This presents the biggest problem when the steam engine is used for transport, since all the fuel needs to be carried along with the engine.

Energy use of a steam train

Let us take as an example a steam engine with a power of 6000 horsepower travelling for 5 hours at full speed. Horsepower is the unit used for measuring engine energy and 1 h.p. is equivalent to 0.75 kW.

The power of the engine is therefore $0.75 \times 6000 = 4500\,\text{kW}$

The engine will use 4500 kJ each second.

So energy used in 5 hours $= 5 \times 60 \times 60 \times 4500 \times 10^3$

$= 81 \times 10^9\,\text{J}$

The energy density of wood is $17\,\text{MJ kg}^{-1}$, so to produce this amount of energy will require $\dfrac{81 \times 10^9}{17 \times 10^6} = 4765\,\text{kg}$ of wood

The energy density of coal is about twice as much as wood so only half the mass of coal would be needed (2223.5 kg)

Coal is also about twice as dense which means the volume of each kg of coal is only half that of wood. The volume of coal needed would therefore be a quarter of the volume of wood. The difference is increased by the fact that wood has a very irregular shape so a pile of wood takes up a lot more space than a pile of coal.

Examiner's hint: To simplify this example the efficiency of the engine has been ignored. In reality, the efficiency of a steam engine was only about 12%. This means that the amount of fuel required was about eight times more than calculated here.

The industrial revolution

The invention of the steam engine led to the possibility of the mechanization of industry and the ability to transport goods and raw materials around the world. This whole mechanization was dependent on coal as a source of energy, and since it was more economical to use the coal near its source, industrial cities grew around areas where coal could be found.

Oil (petroleum)

Oil has also been used for thousands of years but when crude oil comes out of the ground it is a thick sticky substance. Until the development of the technology to drill into the deposits and extract it with pipes, it was more difficult to utilize than coal, even though it does have a higher energy density. In 1852 when Ignacy Łukasiewicz invented a method of refining crude oil to make kerosene (a much cleaner fuel with an even higher energy density of 43.1 MJ kg^{-1}), it became possible to inject the fuel inside the piston of the engine instead of heating it from the outside. The internal combustion engine revolutionized transport; not only did it use fuel with a higher energy density but was about twice as efficient. There is also another advantage of using oil as a fuel – since it is liquid it can be easily pumped from an oil well to the place of use through pipes, with no need for costly transport. However, the fluid nature of oil does lead to one major problem: if a tanker carrying oil spills its load, the oil spreads over a large area. This has led to many environmental disasters over the past 100 years.

Cleaning up after oil spilled from a tanker is washed up on a beach.

Generation of electricity

Until the late 1800s, the main source of household heating in industrialized countries was coal, and the main form of lighting was gas, kerosene or candles. To transport all of these fuels meant using a lot of energy and a lot of time. The way to make electricity from moving a wire in a magnetic field was discovered by Michael Faraday in 1831, but it wasn't until 1866, when Werner Siemens invented the dynamo, that electricity generation on a big scale became a possibility. When in 1884 Sir Charles Pearson invented the steam turbine, all the pieces of the jigsaw could be fitted together. Suddenly, electricity was the easiest way to transfer energy from one place to another.

Exercises

3 When a car is driving at 80 km h^{-1} it is doing work against air resistance at a rate of 40 kW.

(a) How far will the car travel in 1 hour?
(b) How much work does the car do against air resistance in 1 hour?
(c) If the engine of a modern diesel car is 75% efficient, how much energy must the car get from the fuel?
(d) If the energy density of diesel is 45.8 MJ kg^{-1}, how many kg of diesel will the car use?
(e) If the density of diesel is 0.9 kg l^{-1}, how many litres will the car use?
(f) Calculate the litres of fuel used per kilometre.

The coal-fired power station

Cooling towers are a common sight near power stations. They are used to cool down the steam coming out of the turbine. This waste heat could be used to heat the houses nearby. This would increase the efficiency of the power station up to 70%.

Figure 8.10 represents a typical coal-fired power station. The heat from the furnace boils water in the boiler that turns into steam and powers the turbine, the turbine turns a generator and produces electricity. When the steam comes out of the turbine it is cooled, causing it to condense, and this water is then returned to the boiler.

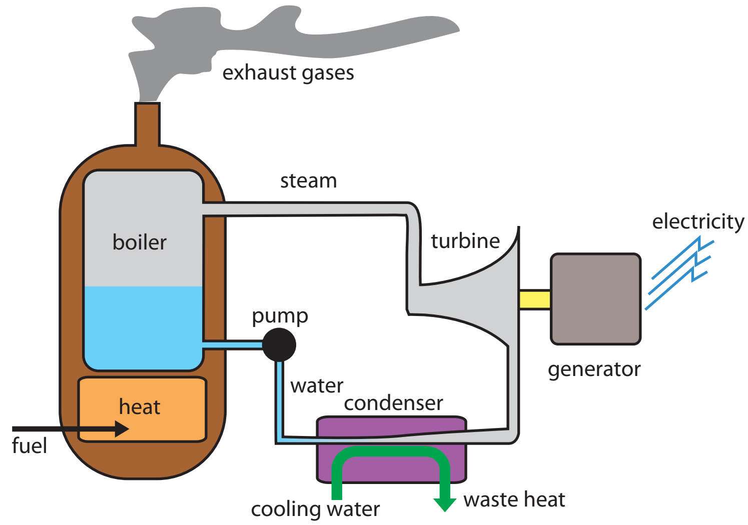


Figure 8.10 In a coal-fired power station the coal is made into dust and blown into a furnace. This produces a lot of smoke that must be cleaned before it is released into the atmosphere.

The overall efficiency of a coal-fired power station is around 40% as not all of the chemical energy from the coal gets converted to electricity. The exhaust gases from the original burning take some of the heat, as does the heat given out when the steam from the turbine condenses. There is also some friction in the components of the turbine and generator.

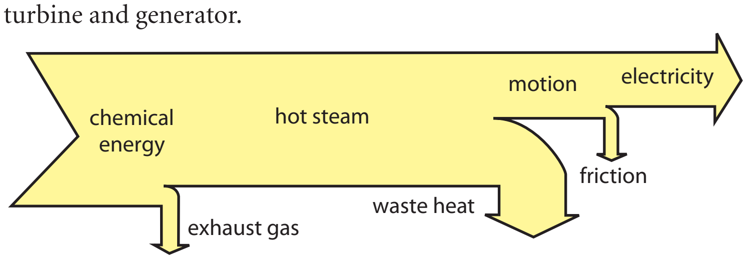


Figure 8.11 Sankey diagram for energy flow in a coal-fired power station.

Oil-fired power station

This is the same as a coal-fired set up but oil is burnt to produce the energy needed to boil the water. Oil is a cleaner fuel than coal and is easier to transport. Using modern drilling technology, it is also much easier to get out of the ground than coal, which still has to be mined.

Exercises

4 A coal-fired power station gives out 1000 MW of power.

(a) How many joules will be produced in one day?

(b) If the efficiency is 40%, how much energy goes in?

(c) The energy density of coal is 32.5 MJ kg^{-1}. How many kg are used?

(d) How many rail trucks containing 100 tonnes each are delivered per day?

Gas-fired power station

Using gas is more efficient than using coal because there can be two stages of energy use. First, the burning gas is blasted through a turbine, then the heat produced by the burning can be used to boil water and power a steam turbine as in Figure 8.13.

Power stations producing electricity from gas can be up to 59% efficient but if the wasted heat is used to heat houses, the overall efficiency can be as high as 80%.

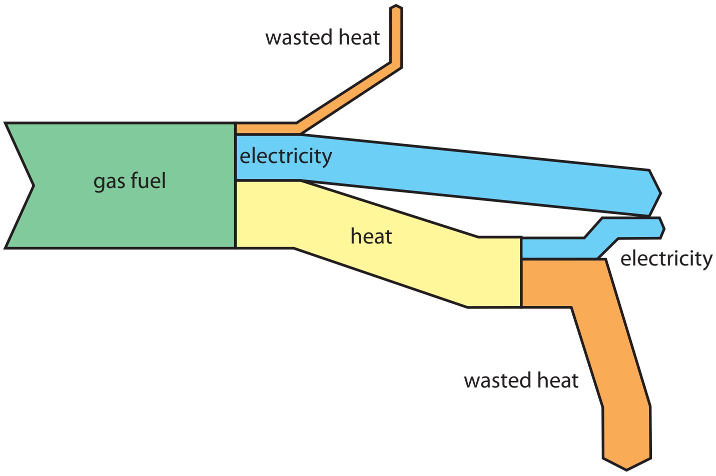


Figure 8.12 Sankey diagram for the flow of energy through a gas-fired power station.

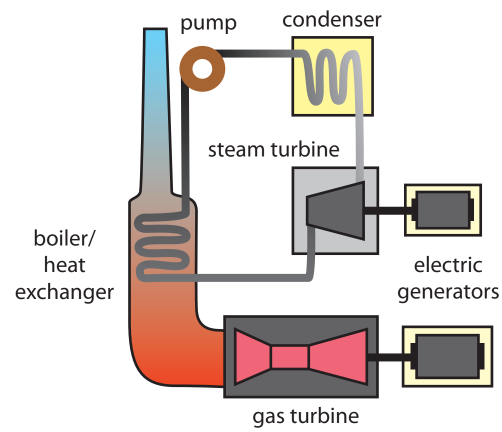


Figure 8.13 The gas-fired power station has two stages of energy use.

8.4 Nuclear power

Assessment statements

8.4.1 Describe how neutrons produced in a fission reaction may be used to initiate further fission reactions (chain reaction).

8.4.2 Distinguish between controlled nuclear fission (power production) and uncontrolled nuclear fission (nuclear weapons).

8.4.3 Describe what is meant by fuel enrichment.

8.4.4 Describe the main energy transformations that take place in a nuclear power station.

8.4.5 Discuss the role of the moderator and the control rods in the production of controlled fission in a thermal fission reactor.

8.4.6 Discuss the role of the heat exchanger in a fission reactor.

8.4.7 Describe how neutron capture by a nucleus of uranium-238 (^{238}U) results in the production of a nucleus of plutonium-239 (^{239}Pu).

8.4.8 Describe the importance of plutonium-239 (^{239}Pu) as a nuclear fuel.

8.4.9 Discuss safety issues and risks associated with the production of nuclear power.

8.4.10 Outline the problems associated with producing nuclear power using nuclear fusion.

8.4.11 Solve problems on the production of nuclear power.

The fission reaction

In Chapter 7 you learnt about *nuclear fission*; this is when a big nucleus such as ^{236}U splits into two smaller nuclei, resulting in a loss of mass and hence a release of energy.

Here is the example:

^{236}U (uranium) splits into ^{92}Kr (krypton) and ^{142}Ba (barium).

Table 2 Nuclear masses of elements with big nuclei.

Z	Symbol	A	Mass (U)
92	U	236	236.045563
56	Ba	142	141.916361
36	Kr	92	91.926270
0	n	1	1.008664

First we can use the data from Table 8.2 and calculate that if this reaction takes place the number of protons balances: $56 + 36 = 92$. But the masses don't balance: $142 + 92 = 234$.

Since the mass of ^{236}U (uranium) was 236 and not 234, two neutrons were emitted in this reaction.

$^{236}U \rightarrow {}^{92}Kr + {}^{142}Ba + 2n$

The amount of energy released can be found from the change of mass:

$236.045\,563 - (141.916361 + 91.926\,270 + (2 \times 1.008\,664)) = 0.185\,604U$

This is equivalent to $931.5 \times 0.185\,604$ MeV $= 172.8$ MeV

To convert this to joules we must remember that 1 eV is the energy gained by an electron accelerated through a potential difference of 1 V.

$E = Vq = 1 \times 1.6 \times 10^{-19}$ J

So, the energy released $= 172.8 \times 10^6 \times 1.6 \times 10^{-19}$ J $= 2.76 \times 10^{-11}$ J.

This energy is given to the fission fragments as KE.

One mole of uranium has a mass of 236 g and contains Avogadro's number of atoms. If a whole mole of uranium split then the energy released would be 16.5×10^{12} J. That is about 63×10^{12} J per kg. This is a lot more energy per kg than coal.

● **Examiner's hint:** This calculation is based on all the nuclei in a pure sample of ^{236}U undergoing fission. The fuel used in a nuclear reactor only contains 3% of ^{235}U which is converted to ^{236}U by absorbing a neutron, and not all of this undergoes fission.

The chain reaction

To split a ^{236}U nucleus requires some energy because the nucleons are held together by a strong force. This energy can be supplied by adding a neutron to a ^{235}U nucleus. This actually increases the binding energy of the nucleus, but because the nucleus cannot get rid of this energy, it splits in two. As a result there are too many neutrons and some are released. These neutrons can be captured by more ^{236}U nuclei, and so on, leading to a chain reaction.

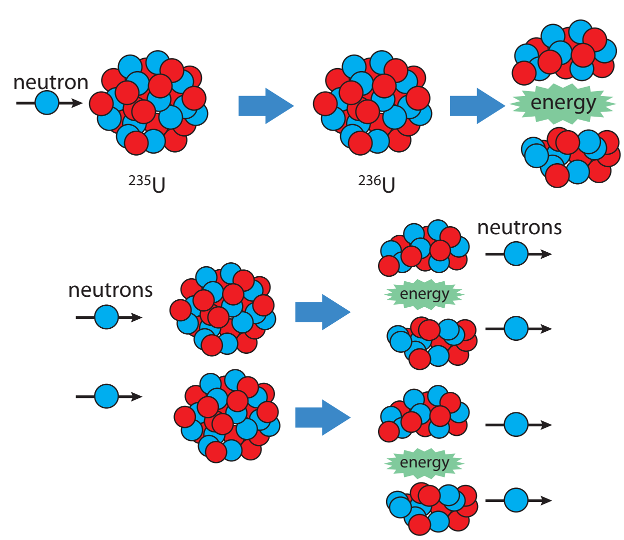


Figure 8.14 A chain reaction from nuclear fission.

Moderation of neutrons

The chain reaction can only occur if the neutrons are moving slowly – otherwise they will pass straight through the nucleus. In fact their KE should be about 1 eV. We calculated earlier that when the nucleus splits it releases about 170 million eV, and although most of this is given to the fission fragments, the neutrons receive a lot more than 1 eV. To make a chain reaction these neutrons need to be slowed down – this is done by introducing some nuclei in between the ^{235}U nuclei. When the neutrons hit these nuclei they lose energy and slow down.

Collisions
When two bodies collide, energy is transferred from one to the other. The maximum transfer of energy occurs if the size of the two bodies is equal. It's not practical to use neutrons as a moderator but in early reactors the deuterium in heavy water was used. This is three times bigger than a neutron but does the job.

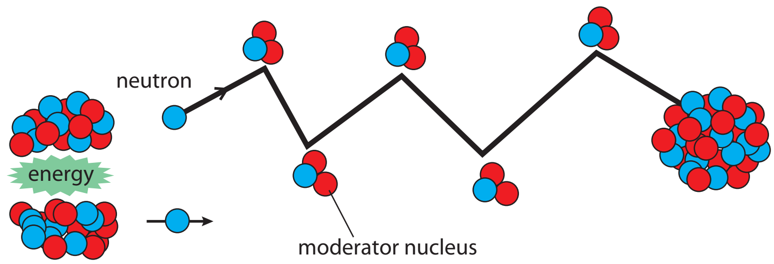


Figure 8.15 A neutron slows down as it collides with moderator nuclei.

Critical mass

Another critical factor that determines whether a chain reaction can take place is the size of the piece of uranium. If it is too small, then before the neutrons have travelled far enough to be slowed down they will have left the reacting piece of uranium. The minimum mass required for a chain reaction is called the *critical mass*.

Nuclear fuel

Uranium is taken out of the ground in the form of uranium ore and processed into a metal that is more than one and a half times as dense as lead. Natural uranium is made up of several different isotopes but 99.3% of it is ^{238}U. This is a weak alpha emitter but is not able to undergo fission; in fact, it absorbs neutrons but doesn't split, so it actually prevents a chain reaction taking place. The isotope ^{235}U is the one useful as a nuclear fuel, but this makes up only 0.7% of the metal. So, before uranium can be used as a fuel the percentage of ^{235}U must be increased in a process called *enrichment*. Uranium used to fuel a nuclear reactor has 3% of the isotope ^{235}U. The ^{238}U that is removed to increase the proportion of remaining ^{235}U is called *depleted uranium*, a heavy, tough metal that is used for making radiation shielding and weapons, particularly armour-piercing ammunition.

For use in a nuclear reactor the fuel is made into small cylinders that are stacked together to make rods (fuel rods). There are many different designs of rod but they are typically 1 cm in diameter and several metres in length. The fuel rods are bundled together making a *fuel bundle*. A reactor will contain many of these bundles.

Pellets of nuclear fuel are stacked into tubes that are bundled together before being put into the reactor.

Plutonium (Pu)

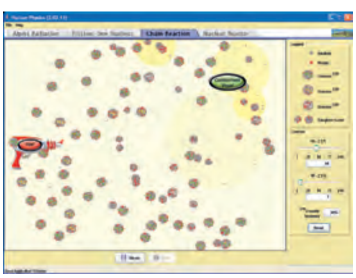
When ^{238}U absorbs a neutron it turns into ^{239}U and this decays by giving out β radiation to form ^{239}Np (neptunium), which in turn decays again by β radiation to ^{239}Pu. Plutonium also undergoes fission and can be used as a fuel or in the manufacture of nuclear weapons. In this way, as the ^{235}U fuel is used up, it actually makes ^{239}Pu. This can be extracted and used for subsequent energy production (or bombs).

Controlling the rate of reaction

The neutrons are the key to making the chain reaction and they are also the key to controlling it. If more than one neutron from each fission goes on to make another fission then the reaction will accelerate; if less than one then it will slow down.

Loss of control: the atom bomb

The atom bomb, or fission bomb, is an out-of-control reaction. This is easy to achieve: if enough ^{235}U is mixed with a moderator, the reaction will go out of control. To stop this happening before the bomb is dropped, the uranium fuel plus moderator is kept in two halves, each half is smaller than the critical mass but when they are put together, a chain reaction starts in the combined piece. The rate of reaction required for an explosion is much greater than for a nuclear reactor, therefore the percentage of ^{235}U in the fuel must be much greater. It is possible to make a bomb with 20% of isotope ^{235}U but 85% is considered 'weapons grade'. The amount of this material required to make an atom bomb would be about the size of a soft drinks can.



The PhET simulation 'Nuclear Physics' shows what happens when the ratio of ^{235}U to ^{236}U is changed as well as showing the effect of the control rods. To try this, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 8.2.

Control in a nuclear reactor

The rate of reaction in a nuclear reactor is limited by the fact that the fuel contains a high proportion of ^{238}U, which absorbs neutrons. Since this cannot easily be altered it cannot be used to slow the reaction down if it goes too fast. This is done by introducing rods of a neutron-absorbing material, such as boron, in between the fuel rods.

Exercises

5 Write the nuclear equation for:

(a) ^{238}U changing to ^{239}U

(b) ^{239}U changing to ^{239}Np

(c) ^{239}Np changing to ^{239}Pu.

The nuclear power station

There are many different designs of nuclear reactor but they all have a nuclear reaction at the core. The energy released when the nuclei split is given to the fission fragments (although about 10 MeV is given to neutrinos that escape). As you know from Chapter 3, the temperature of a body is related to the average KE of the atoms; this means that the temperature of the fuel increases. The hot fuel can then be used to boil water and drive a turbine as in the coal-fired power station.

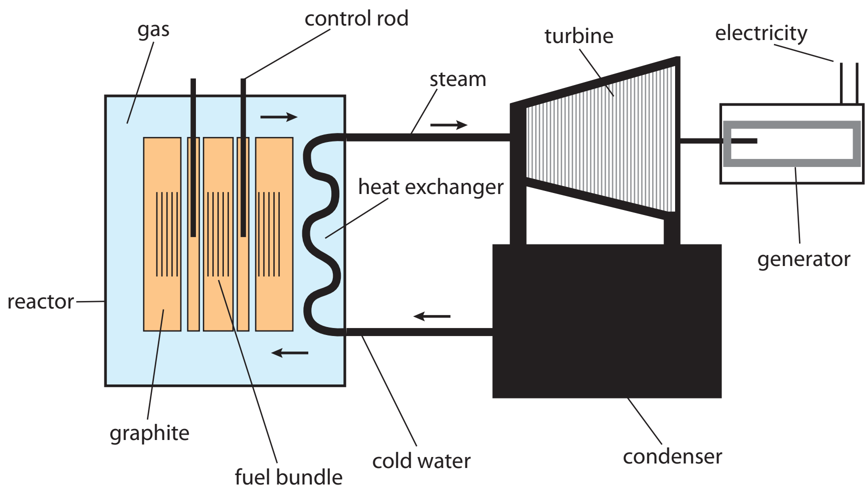


Figure 8.16 An advanced gas-cooled reactor (AGR).

The nuclear reactor is the part that produces heat and contains the fuel rods surrounded by a graphite moderator (yellow). The control rods can be raised and lowered to control the rate of reaction. The nuclear reactor is housed in a pressure vessel in which a gas is circulating (blue). This picks up heat from the fuel rods and transfers it to water in the heat exchanger. This water turns to steam and turns the turbine. The steam cools down and turns back to water in the condenser and is recirculated.

Nuclear submarines
The mechanical energy can also be used directly instead of being used to make electricity. This is particularly useful for powering submarines since the reactor does not use up oxygen and only needs refuelling every 10 years or more.

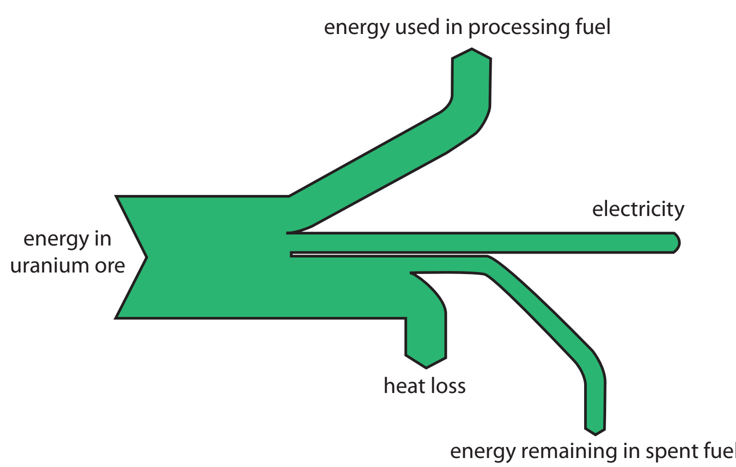


Figure 8.17 Sankey diagram for a nuclear reactor.

Wasted energy
The efficiency of the nuclear reactor is not as high as might be expected. Firstly, the fuel has to be enriched, which takes a lot of energy. Then, it is not possible to get all the energy from the fuel because when the amount of ^{235}U falls below a certain value a chain reaction can no longer be sustained.

Problems with nuclear energy

As with all forms of large scale energy production, there are problems associated with producing power from nuclear energy.

Extraction of uranium

There are several ways of getting uranium out of the ground and none are problem free. Open-cast mining creates environmental damage and underground mining is dangerous for the workers who operate the machines, especially due to the radioactive nature of the rock. Both methods also leave piles of radioactive waste material (tailings). An alternative method involves 'leaching' where solvents are used to dissolve the uranium, which is then pumped out of the ground. However, this can lead to contamination of groundwater.

Meltdown

If the nuclear reaction is not controlled properly it can overheat and the fuel rods can melt: this is called *meltdown*. When this happens the fuel cannot be removed and may cause the pressure vessel to burst, sending radioactive material into the atmosphere. A similar situation occurred in Chernobyl, Ukraine in 1986. However, it is not possible for the reactor to blow up as an atom bomb, since the fuel is not of a high enough grade.

Meltdown can be caused by a malfunction in the cooling system or a leak in the pressure vessel. It would result in severe damage to the reactor, maybe leading to complete shutdown. Further damage outside the reactor is limited by the containment building, an airtight steel construction covered in concrete which not only prevents dangerous material leaking out, but will withstand a missile attack from the outside. Improved reactor design and construction coupled with computer monitoring of possible points of weakness has reduced the possibility of any failure of the structure that might lead to meltdown.

Waste

There are two types of waste associated with nuclear power:

Low level waste

The extraction of uranium from the ground, the process of fuel enrichment and the transfer of heat from the fuel rods all leave some traces of radioactive material that must be carefully disposed of. The amount of radiation given off by this material is not great, but it must be disposed of in places away from human contact for 100–500 years.

Old reactors are another form of low level waste. They cannot simply be knocked down and recycled since most of the parts will have become radioactive. Instead, they must be left untouched for many years before demolition, or they can be encased in concrete.

High level waste

The biggest problem faced by the nuclear power industry is the disposal of spent fuel rods. Some of the isotopes they contain have a half-life of thousands of years so need to be placed in safe storage for a very long time. In the case of plutonium it would not be considered safe for at least 240 000 years. There have been many suggestions: sending it to the Sun; putting it at the bottom of the sea; burying it in the icecap; or dropping it into a very deep hole. For the moment, most of it is dealt with in one of two ways:

- stored under water at the site of the reactor for several years to cool off, then sealed in steel cylinders.
- reprocessed to separate the plutonium and any remaining useful uranium from the fission fragments. This results in waste that is high in concentrations of the very radioactive fission fragments, but the half-life of these fragments is much shorter than either uranium or plutonium, so the need for very long term storage is reduced.

Making weapons from fuel

The fuel used in nuclear reactors does not have enough ^{235}U to be used in the manufacture of atom bombs. However, the same technology used to enrich the uranium to make it into fuel could be used to produce weapons grade uranium. It is, however, plutonium that is the most used isotope in the construction of atomic weapons and this can be obtained by reprocessing the spent fuel rods.

Benefits of fission

No CO_2

The greatest benefit of nuclear power production is that it does not produce any CO_2 or other 'greenhouse gases' and so does not add to global warming.

Sustainability

Although the amount of uranium known to be in the ground would only last about 100 years if used at the same rate as today, the fact that the fission process creates plutonium, which can also be used as a fuel, extends the potential to over 2000 years. Plutonium is, however, a more dangerous fuel to handle than uranium as it is highly radioactive (decaying by alpha emission).

It is also possible to use fuels other than uranium and plutonium.

There seem to be a lot of problems associated with the production of energy from nuclear fuel. However, these problems are not insoluble. Scientists found the way of producing the energy, so surely they can find the solutions to these problems.

Exercises

6 Barium-142 ($^{142}_{56}$Ba) is a possible product of the fission of uranium-236. It decays by β- decay to lanthanum (La) with a half-life of 11months.

(a) Write the equation for the decay of barium.

(b) Estimate how long will it take for the activity of the barium in a sample of radioactive waste to fall to $\frac{1}{1000}$ of its original value.

7 Plutonium-239 splits into zirconium-96 and xenon-136. Use the table to answer the following questions.

Isotope	Mass (U)
^{239}Pu	239.052158
^{96}Zr	95.908275
^{136}Xe	135.907213
neutron	1.008664

(a) How many neutrons will be emitted?

(b) Write the nuclear equation for the reaction.

(c) How much energy is released when the fission takes place?

(d) What is the mass of 1 mole of plutonium?

(e) How many atoms are there in 1 kg of plutonium?

(f) How much energy in eV is released if 1kg of plutonium undergoes fission?

(g) Convert the answer to part **(f)** into joules.

8 A sample of nuclear fuel contains 3% ^{235}U. If the energy density of ^{235}U is 9 x 10^{13} J kg^{-1}, how much energy will 1kg of fuel release?

9 An individual uses around 10 000 kWh of energy in a year.

(a) How many joules is this?

(b) From your answer to question 8, calculate how much nuclear fuel this amounts to.

Fusion

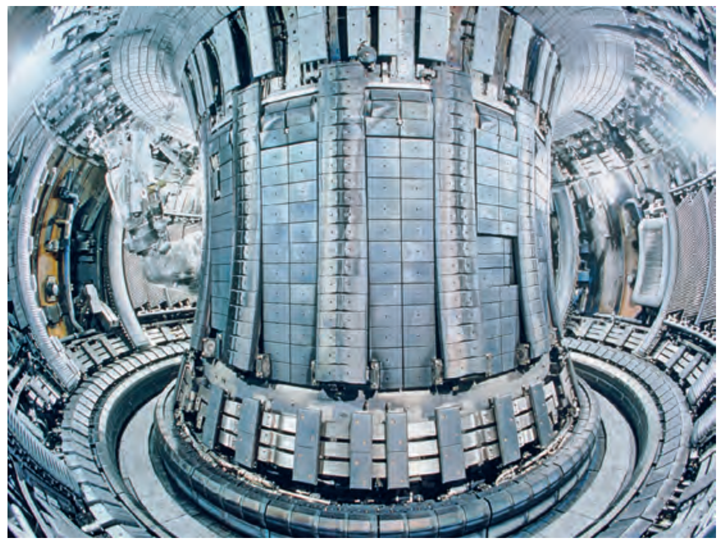
As you learnt in Chapter 7, *fusion* is the fusing together of light nuclei to form larger ones. The larger nuclei have lower mass than the sum and the difference in mass is converted to energy. It certainly works as a way of producing energy – after all this is the way the Sun produces its energy.

In the 1950s fusion was thought to be the energy source of the future and all that had to be done was to solve the technical problems of confinement and sustaining the reaction, then all the energy needs of the world would be satisfied. Fifty years later these problems have still not been solved.

The fusion reactor

Experimental reactors have come very close to producing more energy than the amount of energy put in, although a commercial fusion reactor has yet to be built. The fuel of the fusion reactor is a *plasma*. This is gas in which the electrons and nuclei are separate, but if the plasma is made hot enough, the nuclei can collide with each other fast enough to overcome their electric repulsion and fuse together. The first problem is to find a way to contain plasma with a temperature of 100 million K.

Inside the tokamak the moving charges themselves cause a magnetic field which, together with an external field produced by more magnets, holds the particles away from the walls.



To find more information about the JET project (at the world's largest nuclear fusion laboratory), visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 8.3.

Magnetic confinement

The way that the hot plasma is confined is by using a magnetic field. In Chapter 6 you learnt about how charged particles travel in a circle when moving perpendicular to a magnetic field. By using specially shaped magnets the plasma can be made to travel around the inside of a doughnut-shaped ring – a *tokamak*. Fast-moving particles are difficult to control and the system relies on fast computers to adjust the different magnetic fields.

Heating the plasma

Since plasma contains charged particles, it can be given energy in the same way as giving energy to electrons in a wire. Earlier in this chapter you saw how current is induced in a generator by moving a coil through a magnetic field. In the tokamak instead of moving the wire through the field, the field is changed in such a way as to make the charges move faster around the doughnut.

Burning plasma

When you burn wood on a fire you only need to light it once and the burning wood sets fire to any new wood. It wouldn't be good if you had to light the fire every time a new piece of wood was added; however, this is the problem with fusion reactors so far. The energy comes in bursts: each time new plasma is added, the input of energy needed to raise the temperature sufficiently for the nuclei to fuse is huge.

The fusion bomb

The complete theory of how a fusion bomb works is a military secret. Who owns knowledge?

The fusion bomb, or hydrogen bomb, uses a conventional fission bomb to create enough heat and at the same time compress nuclei so they fuse together. This gives out a huge amount of energy but it is not controllable.

The Sun

The Sun is often described as a burning ball of gas. By measuring the spectrum of the light that is emitted by the Sun we know that it is made of helium and hydrogen. It does not burn in the way that coal burns, but the thermal energy is generated by a fusion reaction. The temperature of the outer part of the Sun, the bit we see, is only 6000 K and not hot enough for fusion to take place. The fusion reaction takes place in the much hotter and denser core (Figure 8.18).

How do we know the temperature of the Sun?
The outer part of the Sun gives out light and its temperature can be calculated from the wavelength of this light. However, this method can't be used for the core, but if we assume that the energy is coming from a fusion reaction, it is possible to calculate the temperature that must be achieved for this reaction to take place.

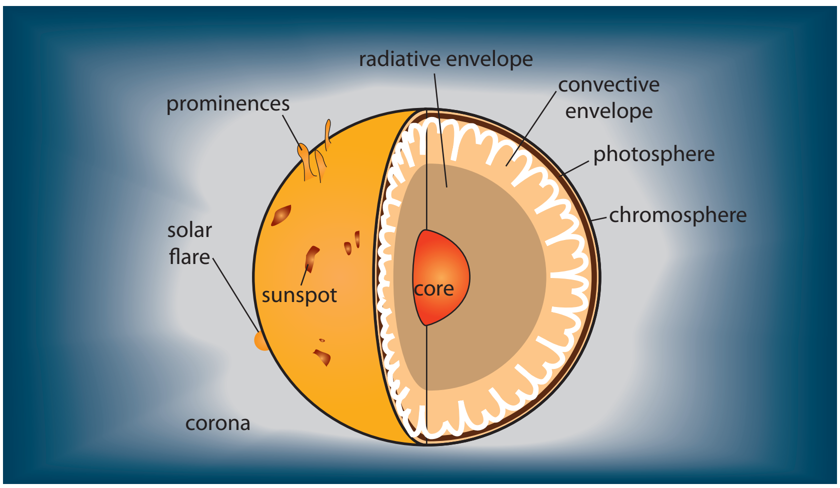
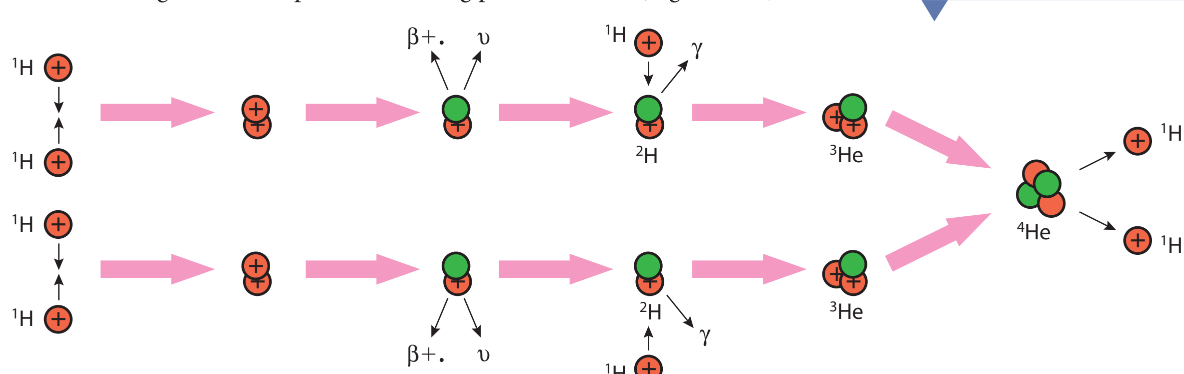


Figure 8.18 The layers and parts of the Sun.

The core

The core of the Sun is a very dense ball of plasma at an estimated temperature of 15 million K (if it wasn't this hot it would collapse due to the gravitational force pulling all the particles together). The fusion reaction is the only energy source that we know of that could produce enough energy to maintain this temperature. It has several stages that take place over a long period of time (Figure 8.19).

Figure 8.19 The energy from the proton-proton chain of the fusion reaction in the core of the Sun radiates through the other layers.



8.5 Solar power

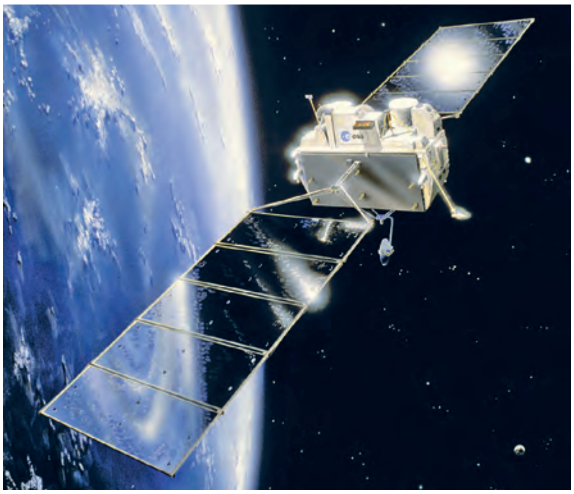
Assessment statements

8.4.12 Distinguish between a photovoltaic cell and a solar heating panel.

8.4.13 Outline reasons for seasonal and regional variations in the solar power incident per unit area of the Earth's surface.

8.4.14 Solve problems involving specific applications of photovoltaic cells and solar heating panels.

Solar panels are used to power satellites. Since these are outside the atmosphere they can gather more energy.



Energy from the Sun

The energy in the form of electromagnetic radiation that is emitted from the Sun in each second is 3.90×10^{26} J. This energy spreads out, and by the time it reaches the Earth, the energy is spread out over a sphere with a radius equal to the Earth's orbital radius of 1.5×10^{11} m. The power per m^2, or intensity, is therefore:

$$\frac{3.90 \times 10^{26}}{4\pi \times (1.5 \times 10^{11})^2} = 1380 \text{ W m}^{-2}$$

This is called the *solar constant*.

The amount that reaches the surface of the Earth depends upon how much atmosphere it has to travel through.

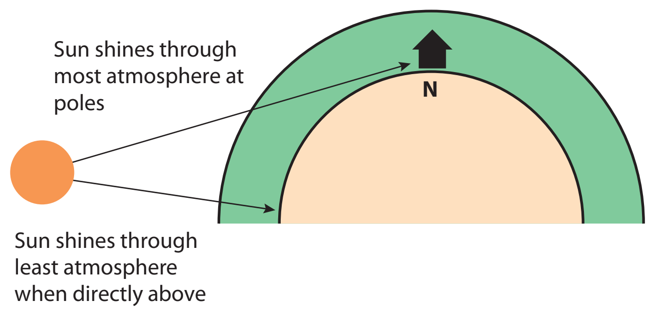


Figure 8.20 Variation of the Sun's intensity with latitude.

Figure 8.20 shows how the intensity varies with latitude. It is also less intense when rising and setting.

When the Sun's radiation lands on the Earth's surface it is either absorbed, causing the surface to get hot, or it is reflected. The amount absorbed depends on the colour of the object. Black/dull objects absorb more radiation than white/shiny ones. So we should use black objects to collect the Sun's radiation.

There are two common ways of using the Sun's energy: either to heat something or to make electricity.

The solar heating panel

A solar heating panel can be used for central heating or for making hot water for household use. They are placed on the roofs of houses (Figure 8.21).

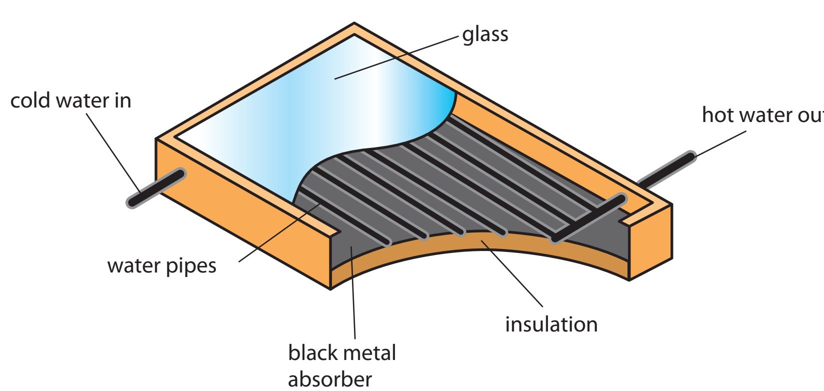


Figure 8.21 A solar heating panel allows radiation from the Sun to enter but prevents the radiation from the warm pipes escaping. This is called the greenhouse effect.

Solar radiation enters the panel through the glass cover, is absorbed by a black metal plate which gets hot and in turn makes the water hot by conduction. Water is continuously circulated so that, as the water gets hot, it flows out and more cold water flows in.

Positioning of solar panels
Solar panels are positioned so that they absorb maximum sunlight in the middle of the day. On the Equator the Sun is directly overhead at midday so the panels are placed horizontally, but in other countries the position of the Sun changes with the seasons, so a compromise has to be made. In countries with bad weather (lots of clouds) the position is not so important because the sunlight does not come from one direction (it is said to be diffuse).

Photovoltaic cell (solar cell)

The photovoltaic cell converts solar radiation into electrical energy. In simple terms, the semiconductors in a photovoltaic cell release electrons when photons of light are absorbed. If different types of semiconductor are placed together then this creates an electrical field that will cause these freed electrons to flow in an external circuit. It's a bit like a battery, but light energy rather than chemical energy is converted to electrical energy. The potential difference (p.d.) and current produced by a single photovoltaic cell is only small, so many cells are usually connected together and can be used to produce power commercially.

Solar heating panels on top of houses in Antalya. The position of solar panels depends on the amount and direction of sunlight as well as the design of the building and other considerations.

Exercises

10 A 4 m^2 solar heating panel is positioned in a place where the intensity of the Sun is 1000 W m^{-2}.
- **(a)** What is the power incident on the panel?
- **(b)** If it is 50% efficient, how much energy is absorbed per second?
- **(c)** If 1 litre (1 kg) of water flows through the system in 1 minute, by how much will its temperature increase? (Specific heat capacity of water = 4 200 J kg^{-1} °C^{-1})

11 A photovoltaic cell of 1 cm^2 is placed in a position where the intensity of the Sun is 1000 W m^{-2}.
- **(a)** If it is 15% efficient, what is the power absorbed?
- **(b)** If the potential difference across the cell is 0.5 V, how much current is produced? (Remember power = IV)
- **(c)** If 10 of these cells were placed in series, what would the total potential difference be?
- **(d)** If 10 of these cells were placed in parallel, what would the current be?
- **(e)** How many of these cells would you need to produce 100 W?

12 Draw a Sankey diagram for a photovoltaic cell.

8.6 Hydroelectric power

Assessment statements
8.4.15 Distinguish between different hydroelectric schemes.
8.4.16 Describe the main energy transformations that take place in hydroelectric schemes.
8.4.17 Solve problems involving hydroelectric schemes.

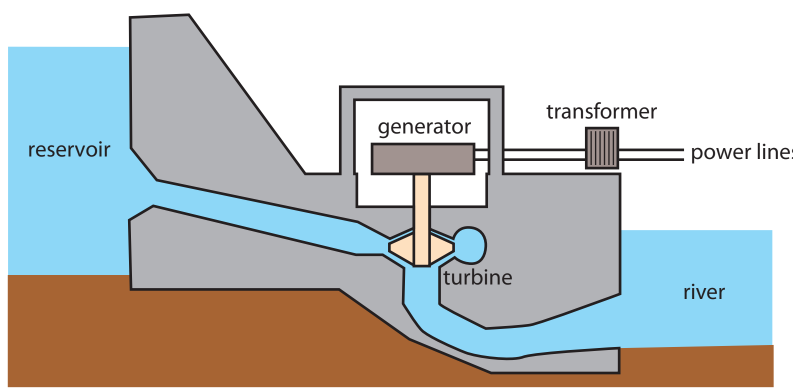
Energy from water

It may not be obvious at first, but the energy converted into electrical energy by hydroelectric power stations comes originally from the Sun. Heat from the Sun turns water into water vapour, forming clouds. The clouds are blown over the land and the water vapour turns back into water as rain falls. Rain water falling on high ground has PE that can be converted into electricity (see Figure 8.22). Some countries like Norway have many natural lakes high in the mountains and the energy can be utilized by simply drilling into the bottom of the lake. In other countries rivers have to be dammed.

The Hoover Dam in Colorado can generate 1.5×10^9 watts.

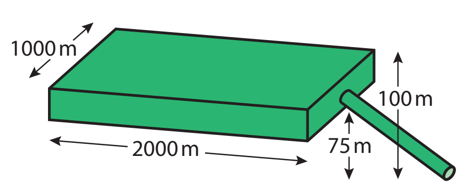
The energy stored in a lake at altitude is gravitational PE. This can be calculated from the equation: PE = mgh where h is the height difference between the outlet from the lake and the turbine. Since not all of the water in the lake is the same height, the average height is used (this is assuming the lake is rectangular in cross section).

Figure 8.22 The main components in a hydroelectric power station.



Worked example

Figure 8.23



Calculate the total energy stored and power generated if water flows from the lake at a rate of 1 m^3 per second.

Solution

The average height above the turbine is

$$\frac{(100 + 75)}{2} = 87.5\,\text{m}$$

Volume of the lake $= 2000 \times 1000 \times 25 = 5 \times 10^7\,\text{m}^3$

Mass of the lake = volume × density $= 5 \times 10^7 \times 1000$

$= 5 \times 10^{10}\,\text{kg}$

PE $= mgh = 5 \times 10^{10} \times 9.8 \times 87.5 = \mathbf{4.29 \times 10^{13}\,J}$

If the water flows at a rate of 1 m^3 per second then 1000 kg falls 87.5 m per second

So the energy lost by the water $= 1000 \times 9.8 \times 87.5 = 875\,000\,\text{J s}^{-1}$

Power = 875 kW

Pumped storage schemes

Most countries produce electricity from a variety of sources such as burning fossil fuel, nuclear power and hydroelectric power. At night when demand is low, it is possible to turn off the hydroelectric power, but if you put out the fire in a coal-fired station it takes a long time to get hot again. The excess power produced from coal-fired power stations can be used to pump water up into a reservoir which can be used to drive the turbines in the daytime. In the long run, this reduces the amount of fossil fuel that needs to be burnt.

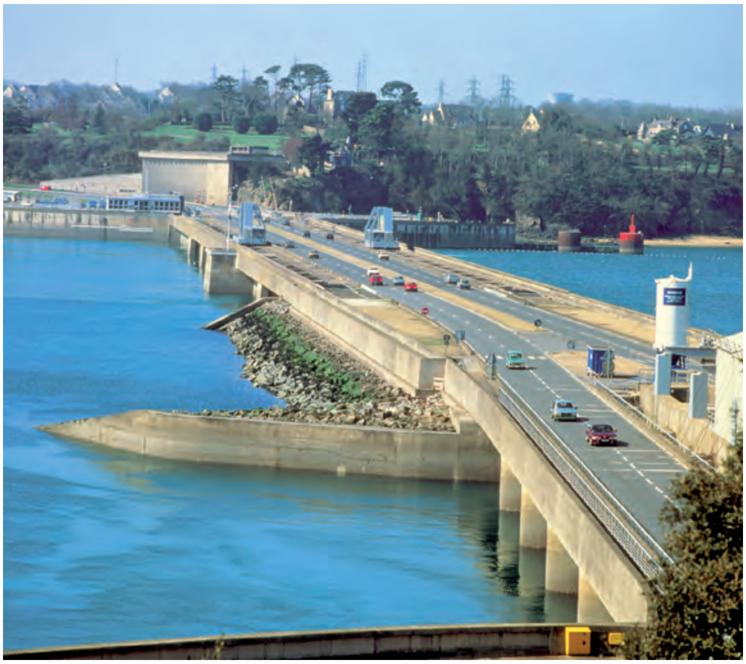
Run-of-the-river power stations

One of the big problems with hydroelectric power in countries without mountain lakes is the need to dam river valleys to create a difference in height to drive the turbines. Run-of-the-river power stations use water that has been diverted from a fast-flowing river without damming the river.

Small is good

Whenever electrical energy is produced it must be transmitted through wires to the place where it will be used. However, passing current through wires results in energy loss because the wires get hot. To reduce this loss, it is quite common for factories that are dependent on large amounts of electrical power, e.g. the production of aluminium, to be sited next to power stations. An alternative approach is to build small-scale power stations near to where people live.

The tidal barrage at Rance in France. After water has flowed into the estuary on the incoming sea tide it is held back and then released through 24 turbines producing 240 MW. The tidal range is up to 13.5 m.



Tidal power

Where there is a big difference between high and low tide, the tidal flow can be used to drive turbines and produce hydroelectric power. One way to do this is to build a dam or barrage across a river estuary. The water is held back as it flows in and out of the estuary and then released to drive turbines. An alternative is to fix turbines to the bottom of the estuary and allow the free flowing water to turn them.

Exercises

13 The Hoover Dam is 221 m high with an area of 694 km^2.
  (a) Estimate the mass of water in the dam.
  (b) How much PE is stored in this water?
  (c) The Hoover Dam can produce 1.5×10^9 W of electricity. If the power station is 80% efficient how much PE must be lost per second from the water to produce this power?
  (d) What mass of water must flow though the turbines each second to produce this power?

14 A mountain hut has 4×50 W light bulbs, a 1 kW electric heater and a 2 kW cooker.
  (a) Calculate the total power consumed if all appliances are in use.
  (b) Outside the hut is a 5 m high waterfall. The owner is interested in building a small hydro generator. How many kg of water must flow per second to generate enough power? (Assume 100% efficiency.)

8.7 Wind power

Assessment statements
8.4.18 Outline the basic features of a wind generator.
8.4.19 Determine the power that may be delivered by a wind generator, assuming that the wind KE is completely converted into mechanical KE, and explain why this is impossible.
8.4.20 Solve problems involving wind power.

Wind

Here we have only touched on the basic physical principles behind wind production. The complete picture is far more complicated. In fact, to apply our kinetic model of a gas to the movements of air around the world would be rather pointless. In these situations a systems approach is much simpler.

Using the wind as a source of energy is nothing new; it is over a thousand years since the first known use in Persia was recorded. In those days windmills were used to grind (or mill) corn (hence the name); now they are used to generate electricity. The energy in the wind originates from the Sun. In simple terms the Sun heats the air which becomes less dense and rises, leaving an area of low pressure close to the Earth. Surrounding air will move into this low pressure area and this air movement is the wind. The rotation of the Earth causes this moving air to move in a circular pattern, causing the weather systems that we are familiar with.

Coastal winds

Coastal areas are particularly windy due to the different rates of heating of the land and the sea. During the day, when the Sun is shining, the land and sea absorb energy and get hot. The sea has a bigger specific heat capacity than the land so the temperature of the sea does not rise by as much as the temperature of the land ($Q = mc\Delta T$). The result is that the air above the land rises and this causes a low pressure that allows the air above the sea to flow in. At night the reverse happens, when the land cools down more quickly than the sea.

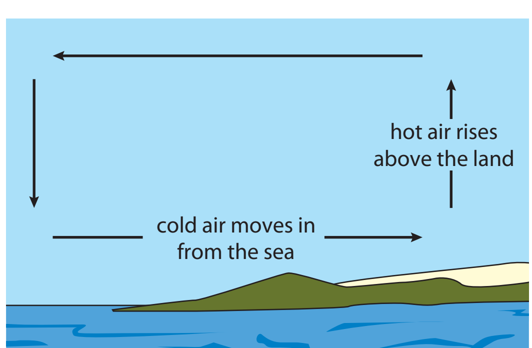


Figure 8.24 Flow of coastal winds during the day

Katabatic winds

A katabatic wind is formed when a high pressure is caused by dense cold air pressing down at the top of a mountain, resulting in air flowing downhill. An example of this regularly takes place when cold air from the Alps and Massif Central areas in France descend towards the Mediterranean coast. Funnelling by the Rhone valley causes the air to speed up as it reaches the sea, causing a strong wind called the Mistral.

A wind turbine with the generator on top of the tower in order to reduce the number of moving parts and hence energy loss due to friction.

The wind turbine

Wind turbines are rather like a fan or the propeller of an aeroplane, except they are moved by the air rather than making the air move. These large turbines are often grouped together in wind farms.

Energy calculation for a wind turbine

Wind has energy in the form of KE. This enables the wind to do work against the turbine which turns a generator creating electrical energy. To calculate how much energy there is in the wind, we consider a cylinder of air with a radius the same as the radius of the turbine as shown in Figure 8.25.

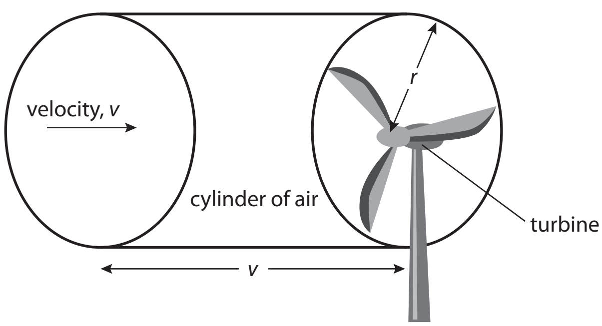


Figure 8.25 Energy from air approaching a wind turbine

If the velocity of air is v then in 1 s it will move a distance v. The volume of air passing by the turbine per second is therefore $v \times \pi r^2$ where r is the length of one of the turbine blades.

The mass of this cylinder of air, $m = \rho v \pi r^2$ where ρ is the density of air.

The KE of this air $= \frac{1}{2}mv^2 = \frac{1}{2}\rho v \pi r^2 v^2 = \frac{1}{2}\rho \pi r^2 v^3$.

Since this is the KE of air moving past the turbine per second it gives us the power in the wind.

The wind doesn't stop after passing the turbine so not all of this energy is turned into electricity. The maximum theoretical percentage of the wind's energy that can be extracted using a turbine is 59%.

Windy places

The best place to put a wind turbine is obviously in a windy place. However, wind speed isn't the only consideration. It is also important that the wind is fairly regular so that the turbine doesn't have to keep changing its orientation. Another factor is how easy it is to lay power lines to the turbine and how easy it is to build the turbine in that position. The main problems associated with wind power are that the turbines often need to be built in areas of natural beauty and that they are unreliable – when the wind stops they produce no electricity.

W

Wind power in Denmark

Denmark produces 20% of its electricity from wind power. This is because Denmark has a long flat windy coastline that is ideal for building wind farms.

To find more information about wind power in Denmark, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 8.4.

Exercises

15 A turbine with a turbine blade length of 54 m is operated in a wind of speed $10\,\mathrm{m\,s^{-1}}$. The density of air is $1.2\,\mathrm{kg\,m^{-3}}$.

(a) How much power is in the wind passing through the turbine?

(b) How much electrical power can be generated if the turbine is 20% efficient?

(c) If the wind speed increased to $15\,\mathrm{m\,s^{-1}}$, how much power would be produced?

8.8 Wave power

Assessment statements

8.4.21 Describe the principle of operation of an oscillating water column (OWC) ocean-wave energy converter.

8.4.22 Determine the power per unit length of a wavefront, assuming a rectangular profile for the wave.

8.4.23 Solve problems involving wave power.

Waves

If you have ever watched waves crashing into a beach on a stormy day you will have realised that there is a lot of energy transmitted in water waves. Waves in the sea are caused by winds disturbing the surface of the water; these winds can be local, in which case the waves tend to be small and with a short wavelength. The big powerful rolling waves favoured by surfers originate way out in the deep ocean. The weather map in Figure 8.26 shows the typical situation that would cause big waves to arrive at the surfing beaches of western Europe.

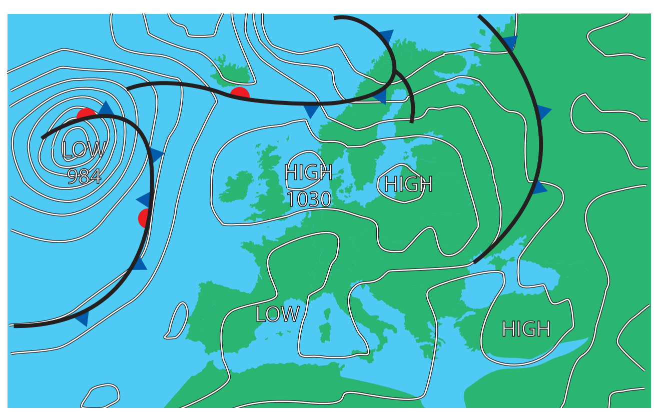


Figure 8.26 In this weather map the low pressure system in the middle of the Atlantic is a storm. This will create waves that travel towards Europe. The amount of energy carried by the waves is directly related to the duration of the storm.

Power in a water wave

Calculating the power in a wave is quite difficult but we can make an estimate if we simplify the situation. The energy in a wave alternates between PE as the water is lifted up, and KE as it falls.

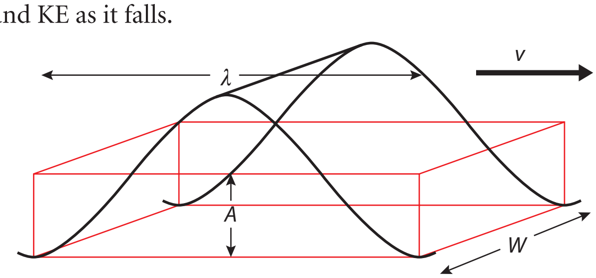


Figure 8.27 The form of a wave can be approximated to a rectangle (length λ, height A and width W travelling at velocity v).

The PE of this mass of water is given by PE $= mgh$

where $h =$ the average height of the wave $= \dfrac{A}{2}$

so PE $= \dfrac{mg\,A}{2}$

But if the density of water $= \rho$

then $m = \rho \times$ volume $= \rho \times \lambda AW$

so PE $= \dfrac{\rho\lambda AWgA}{2} = \dfrac{\rho\lambda WgA^2}{2}$

Power = energy per unit time, so if the waves arrive every T seconds then

power $= \dfrac{\rho\lambda WgA^2}{2T}$

But $\dfrac{\lambda}{T} =$ wave velocity, v

so power $= \dfrac{\rho vWgA^2}{2}$

The power per unit length of wavefront is therefore $\dfrac{\rho vgA^2}{2}$

Generating electricity from water waves

The oscillating water column

The principle of the oscillating water column is shown in Figure 8.28 and consists of a column that is half full of water, such that when a wave approaches it pushes water up the column. This compresses the air that occupies the top half, pushing it through a turbine which drives an electric generator. The turbine is specially designed so that it also turns when the water drops back down the column, pulling air into the chamber.

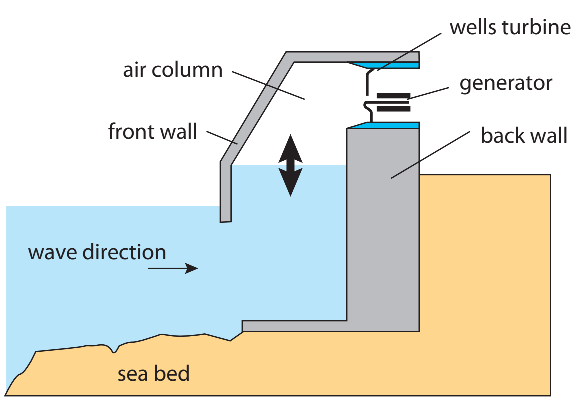


Figure 8.28 The main components of an oscillating water column generator.

Pelamis

The pelamis is named after a sea snake because that's what it looks like. Each pelamis is made of four sections with a total length of 150 m. Each section is hinged, and when a wave passes it bends. The bending drives pumps that move fluid back and forth, powering electrical generators. A pelamis can generate 750 kW of electricity.

W To find out more about wave power, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 8.5.

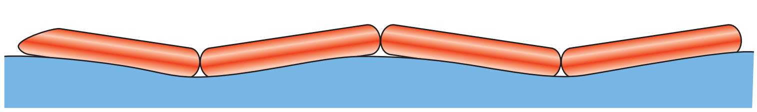


Figure 8.29 The four sections of a single pelamis.

Exercises

16 Waves of amplitude of 1 m roll onto a beach at a rate of one every 12 s. If the wavelength of the waves is 120 m, calculate

(a) the velocity of the waves

(b) how much power there is per metre along the shore

(c) the power along a 2 km length of beach.

8.9 The greenhouse effect

Assessment statements

8.5.1 Calculate the intensity of the Sun's radiation incident on a planet.

8.5.2 Define *albedo*.

8.5.3 State factors that determine a planet's albedo.

8.5.4 Describe the greenhouse effect.

8.5.5 Identify the main greenhouse gases and their sources.

8.5.6 Explain the molecular mechanisms by which greenhouse gases absorb infrared radiation.

8.5.7 Analyse absorption graphs to compare the relative effects of different greenhouse gases.

8.5.8 Outline the nature of black–body radiation.

8.5.9 Draw and annotate a graph of the emission spectra of black bodies at different temperatures.

8.5.10 State the Stefan–Boltzmann law and apply it to compare emission rates from different surfaces.

8.5.11 Apply the concept of emissivity to compare the emission rates from the different surfaces.

8.5.12 Define surface heat capacity C_s.

8.5.13 Solve problems on the greenhouse effect and the heating of planets using a simple energy balance climate model.

The greenhouse effect is the warming of a planet due its atmosphere allowing in ultraviolet radiation from the Sun, but trapping the infrared radiation emitted by the warm Earth. This is similar to the way that the glass of a greenhouse warms the plants inside, hence the name. There are many physical principles that need to be explained before we can fully understand how this effect works: firstly, how the energy from the Sun gets to the Earth; and secondly, the interaction between this energy and the atmosphere.

Solar radiation

As explained in an earlier section, the Sun radiates 3.9×10^{26} joules per second. This energy spreads out in a sphere. By the time it reaches the Earth at a distance of 1.5×10^{11} m, the power per m^2 (or intensity) is $\dfrac{3.9 \times 10^{26}}{4\pi \times (1.5 \times 10^{11})^2} = 1380\ \mathrm{W\,m^{-2}}$

The intensity of radiation at each planet is related to its distance from the Sun: Mercury, the closest planet, receives more power per m^2 than Neptune, the furthest.

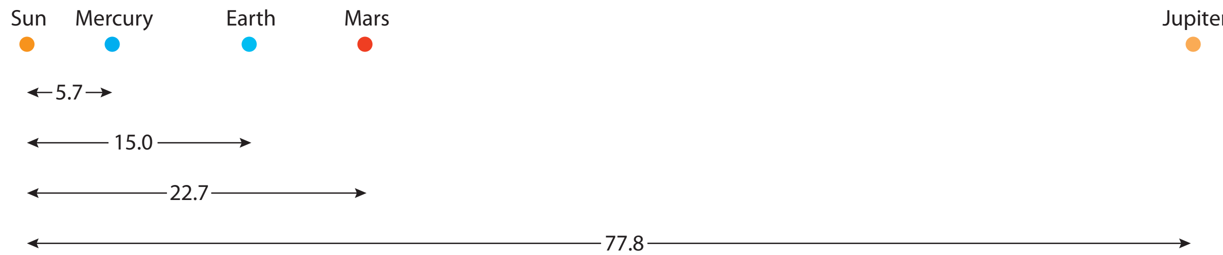


Figure 8.30 The solar system with planet orbits drawn to scale. The size of the planets and the Sun are not to scale – they would be much smaller. The units are 10^{10} m.

Exercises

17 Calculate the intensity of the Sun's radiation on the surface of
(a) Mercury
(b) Jupiter.

Interaction between light and matter

Exciting electrons

The light that comes from the Sun is made up of photons of many different wavelengths. When a photon interacts with an atom, it can give energy to the atom by exciting one of its electrons into a higher energy level, as explained in Chapter 7. This can only happen if the energy of the photon is exactly the same as the energy needed to excite the electron, and when it happens the photon disappears. This can be reversed, and an electron that has been excited into a higher energy level can give a photon back out again, when the electron goes back to its original energy level. This is what happens when light is scattered by the atmosphere.

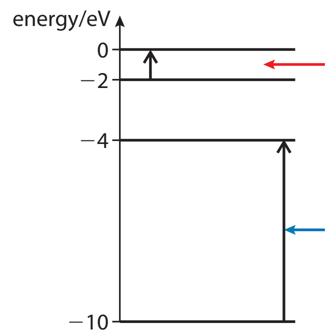


Figure 8.31 Different wavelengths excite electrons to different energy levels.

Ionization

If the photon is of high frequency and therefore high energy (remember $E = hf$) then an absorbed photon can cause an electron to be ejected from an atom. This is called *ionization*. When this happens the energy of the photon doesn't have to be an exact value, just big enough to get the electron away from the atom.

Excitation of molecules

In previous chapters we have mainly been interested in individual atoms rather than molecules. However, to understand the greenhouse effect we must look at how molecules can absorb a photon. A molecule is made of several atoms held together by the electromagnetic force. A simple model would be two balls joined together with a spring as in Figure 8.32. According to what you have learnt in this course, a single ball can only have three types of energy, kinetic due to its movement, potential due to its position and, thirdly, internal energy. If you consider the balls in the figure, they can also be made to oscillate or rotate. If the frequency of oscillation of a molecule is the same as the frequency of a photon, then the molecule can absorb the photon. This causes the molecules to move more and hence have a higher temperature. This is an example of resonance. The frequencies that cause this sort of vibration tend to be quite low, in the infrared region of the spectrum.

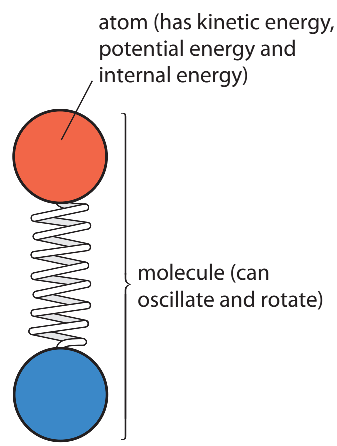
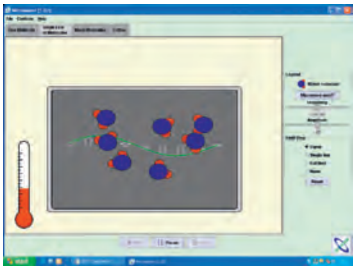


Figure 8.32 A simple model of atoms in a molecule.



The PhET simulation 'Microwaves' shows how water molecules are excited by microwave radiation. To view, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 8.6.

Interaction with solids

When dealing with electron energy levels in Chapter 7, we were considering individual atoms. However, a solid is made up of many atoms that interact with each other. When this happens the electrons no longer have to exist in special (discrete) energy levels but they can have many different energies, and these different energies form bands. This means that solids can absorb many different wavelengths of light rather than just a few special ones. Another consequence of this structure is that it is easier for the energy absorbed by the electrons to be passed on to the atoms, resulting in an increased temperature. When the molecules of a solid are given energy, they can vibrate and give out low frequency radiation. So when light is absorbed, it causes a solid to get hot, resulting in the emission of infrared radiation.

Albedo

When electromagnetic radiation is incident on a surface it is either absorbed, causing the surface to get hot, or it is reflected. The ratio of reflected to incident radiation is called the *albedo*. The albedo for snow is high (90%) since it reflects most of the radiation incident on it, whereas a dark forest has a low albedo of around 10%. The average for the planet Earth including its atmosphere is 30%.

Black body radiation

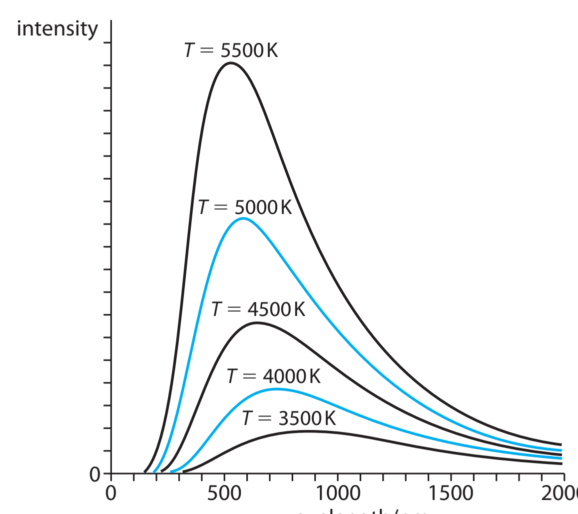
Figure 8.33 A continuous spectrum from a solid and the line spectrum for the gas hydrogen.

As mentioned previously, due to their atomic structure, solids can absorb many different wavelengths of radiation: for the same reason, if a solid is heated it will emit a wide range of frequencies. If we observe the spectrum of light from a solid we see it is continuous because it is made up of many wavelengths. The spectrum from a low pressure gas, on the other hand, consists of just a few lines as in Figure 8.33.

A black object is an object that absorbs all wavelengths, and if heated, it will emit all wavelengths, too. However, not all frequencies will be equally intense. The spectrum for light emitted from a black body is shown in Figure 8.34. The peak of this graph represents the most intense part of the spectrum; this is dependent on the temperature of the body (T). The hotter it is the shorter wavelength this will be. This wavelength (λ_{max}) can be calculated from Wien's displacement law:

$$\lambda_{max} = \frac{b}{T} \text{ where } b = 2.89 \times 10^{-3}\ \text{mK}$$

Figure 8.34 The intensity distribution for a black body at different temperatures.

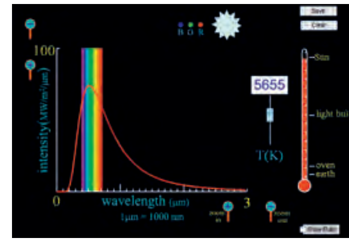


Red hot

When a rod of metal is heated to around 1000 K it starts to glow red. Although the most intense part of the spectrum is not in the visible region, there is enough visible red light to make the rod glow.

Stefan–Boltzmann Law

From the graph in Figure 8.34 we can see that as the temperature of a black body increases, the intensity of the radiation also increases. In other words, the amount of energy emitted from the surface increases. The Stefan–Boltzmann Law relates the power emitted per unit area to the temperature of the surface with the equation:
Power per unit area $= \sigma T^4$ where $\sigma = 5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$



W The PhET simulation, 'Blackbody', will help you to see how the temperature of a body affects the radiation emitted. To view this, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 8.7.

Exercises

18 Calculate the amount of energy in one photon of light of wavelength 600 nm.
The Sun is a hot dense gas so can be treated like a black body.

19 The temperature of surface of the Sun is 6000 K. Use Wien's displacement law to calculate the most intense wavelength in the spectrum of emitted radiation.

20 Use the Stefan–Boltzmann Law to calculate the power per $\mathrm{m^2}$ emitted from the Sun.

21 The Sun's radius is 7×10^8 m. Use your answer to Question 20 to calculate the total energy radiated by the Sun per second.

Radiation from the Sun

As calculated in Question 19, the spectrum of electromagnetic radiation from the Sun has a maximum intensity around 480 nm. This is the wavelength of blue light as seen in the spectrum shown in Figure 8.33. This means that the visible part of the spectrum is right on the peak of the intensity curve – it is no surprise therefore that this is the part we have evolved to see.

Temperature of the Earth with no atmosphere

Energy comes from the Sun to the Earth; this warms up the Earth and the Earth radiates some energy back into space. As the temperature of the Earth increases, it will radiate more and more until the amount of energy radiated = amount of energy absorbed, at which point the temperature will stay constant.

If we ignore the atmosphere we can calculate how hot the Earth should be. We know that the energy incident on the Earth per second is $1360\,\mathrm{W\,m^{-2}}$. As the Earth only catches radiation on one side at a time, we can simplify this as a disc of radius 6400 km. The energy incident on the whole Earth per second is

$$1360 \times \pi \times (6.4 \times 10^6)^2 = 1.75 \times 10^{17}\,\mathrm{W}$$

But the average albedo of the Earth is 30%. This means that 70% is absorbed, therefore:

$$\text{absorbed energy per second} = 1.23 \times 10^{17}\,\mathrm{W}$$

To calculate the heat radiated by the planet, we have to consider the complete area of the sphere: if the temperature is T then (using Stefan–Boltzmann) the radiated energy per second is:

$$(5.67 \times 10^{-8} \times T^4) \times 4\pi r^2 = 2.9 \times 10^7 \times T^4$$

When this equals the heat absorbed, the temperature will stay constant so:

$$2.9 \times 10^7 \times T^4 = 1.23 \times 10^{17}$$

This gives a value of $T = 255\,\mathrm{K}$ ($-18\,°\mathrm{C}$)

This is quite a lot colder than the Earth actually is, because the Earth's atmosphere absorbs some of the radiation radiated by the Earth. This is the greenhouse effect.

Emmisivity (*e*)

The reason that the temperature in the previous example was too low was that the Earth does not radiate as well as a black body.

The emissivity, *e*, of the Earth tells us how it compares to a black body; it is the ratio of the energy radiated by a body to the energy radiated by a black body of the same temperature.

The average temperature of the Earth is 288 K. From the previous calculation we know that if the Earth was a black body it would radiate

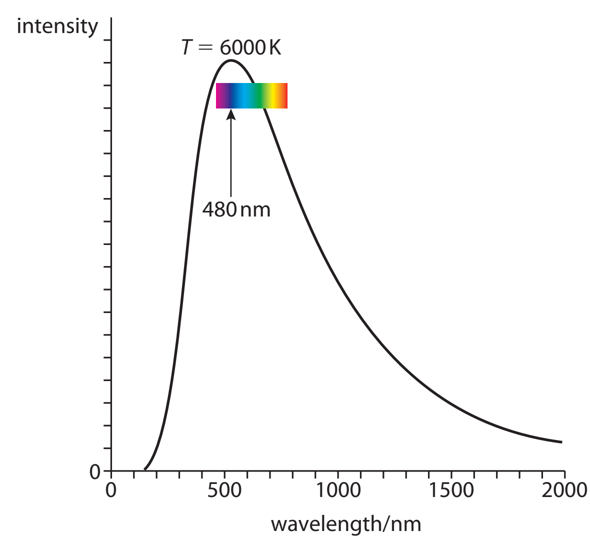
$$2.9 \times 10^7 \times 288^4\,\text{W} = 1.99 \times 10^{17}\,\text{W}$$

But we know that since the temperature is stable the Earth must be radiating 1.23×10^{17} W (the same as the radiation from the Sun) so the emissivity is $\frac{1.23}{1.99} = 0.6$

Absorption by the atmosphere

Before the radiation from the Sun lands on the surface of the Earth it first has to pass through the atmosphere. The atmosphere is in layers, the outer layers containing ozone (a form of oxygen whose molecules are made of three oxygen atoms, O_3). This absorbs the higher-energy parts of the spectrum, ultraviolet and X-rays. If it were not for this layer, these harmful rays would reach the Earth. As the radiation passes through the lower layers, infrared radiation is absorbed by water vapour and carbon dioxide. These gases have molecules that can be excited by the frequency equal to that of infrared photons.

Figure 8.35 The solar spectrum showing all wavelengths emitted by the Sun.



Absorption spectroscopy

The amount of different wavelengths absorbed by each gas can be found by measuring the intensity of light of known wavelength passing into and out of the gas. Figure 8.36 shows the absorption spectra for three different gases showing how ozone absorbs UV, and how carbon dioxide and water absorb IR. Notice that none of these gases absorbs visible light.

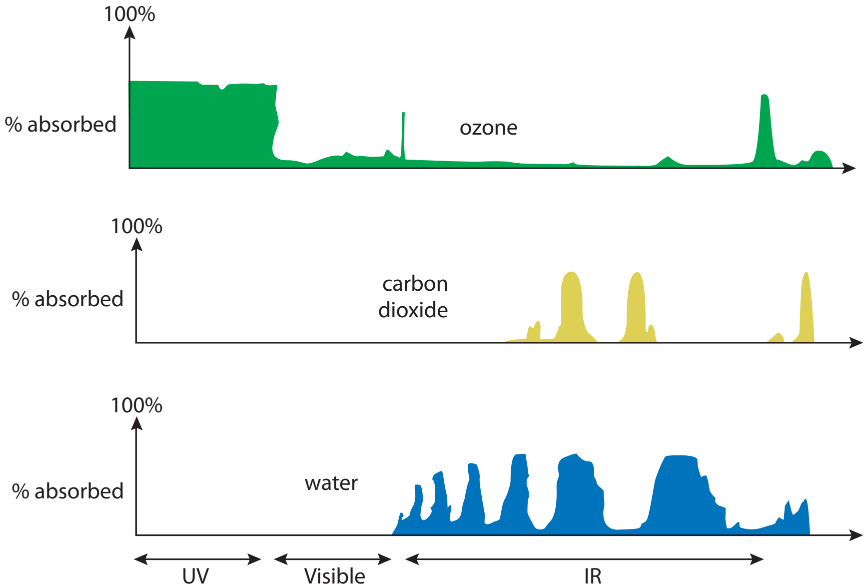


Figure 8.36 Absorption spectra for ozone, carbon dioxide and water.

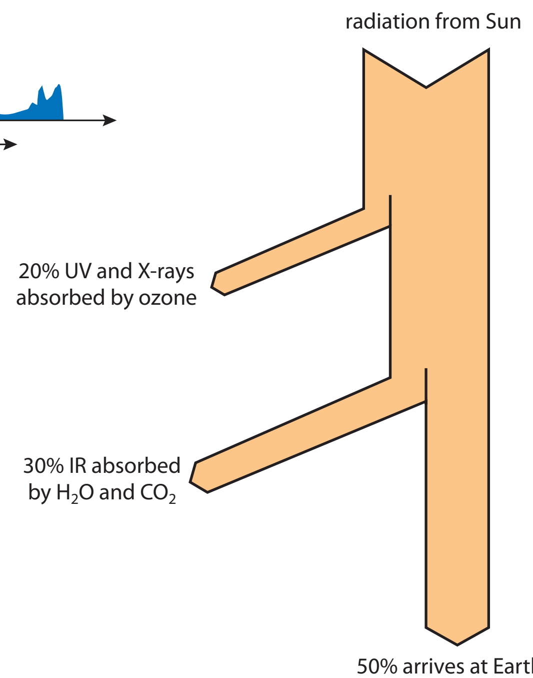


Figure 8.37 Sankey diagram showing solar radiation absorbed by the atmosphere.

Absorption by the ground

When solar radiation lands on the ground, a fraction, depending on the albedo, is reflected straight back. Since this radiation is the same as the radiation that has just made it through the atmosphere, it goes back through the atmosphere without being absorbed. The remaining radiation (mostly visible light) is absorbed and this increases the temperature of the ground. Since the ground is not very hot, then according to Wien's law, the wavelength of the emitted radiation will be in the IR region.

Surface heat capacity

The temperature increase of the ground can be calculated from the surface heat capacity (C_s). This is the amount of heat required to raise the temperature of 1 m^2 of the ground by 1 K. For the Earth this is $4 \times 10^8\ J\ km^{-2}$.

To see how the greenhouse gases contribute to the warming of planet Earth, visit www.heinemann.co.uk/hotlinks. enter the express code 4426P and click on Weblink 8.8.

The greenhouse effect

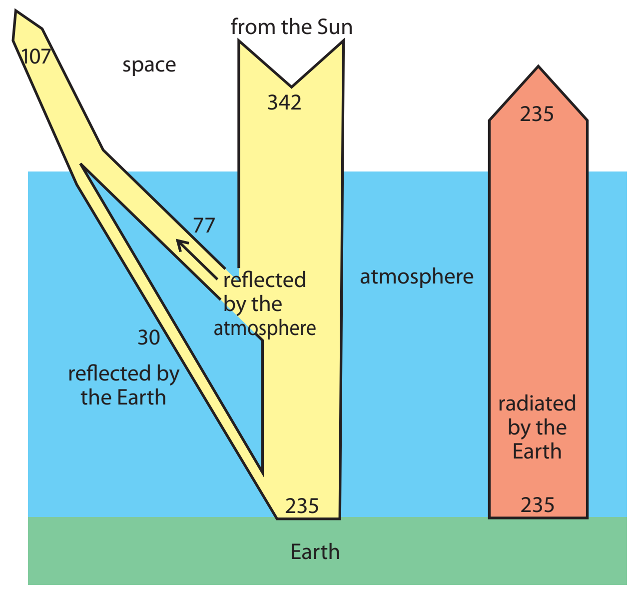
The IR radiation radiated from the ground travels upwards through the atmosphere and as it does it is absorbed by CO_2 and H_2O. As a result, these molecules become excited, raising the temperature of the atmosphere. They themselves can then radiate IR radiation in all directions, some of which goes back to the Earth. We have seen that the temperature of the Earth is determined by the point at which the radiation leaving the Earth is equal to that arriving. By reducing the amount leaving, the temperature at which this balance will be achieved will be higher.

Sankey diagrams

We have calculated already that by the time it gets to the Earth the intensity of the Sun's radiation is 1380 W m^{-2}. However, the Earth is a sphere, so the power received per unit area will be different at different places (less at the poles than at the equator). The average value is taken here to be 342 W m^{-2}.

To understand the energy flow diagram for the complete greenhouse effect let us first consider the flow diagram with no greenhouse gases (Figure 8.38).

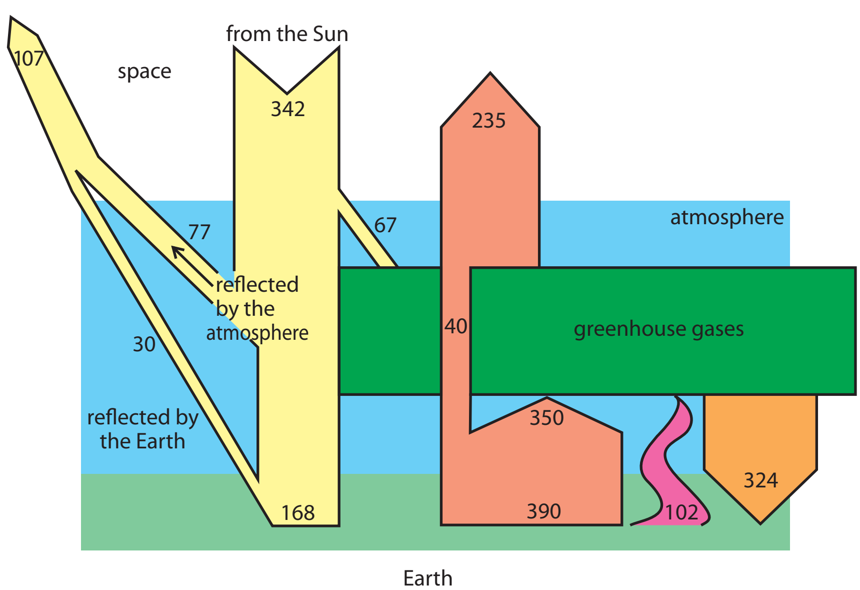
Figure 8.38 Sankey diagram for energy flow without greenhouse gases. The amount of radiation absorbed by the ground equals the amount being radiated.



Energy flow with greenhouse gases
Here we can see that 342 W m^{-2} still enter and leave but now the Earth is radiating more energy. This energy is recirculated by the greenhouse gases. 519 W m^{-2} are absorbed by the gases and 519 W m^{-2} are given out while 324 W m^{-2} of this do not leave the atmosphere.

But we know that due to the greenhouse effect, the Earth is hotter than in this model and so emits more radiation.

Figure 8.39 Sankey diagram for energy flow for the Earth and greenhouse gases in the atmosphere. The 102 W m^{-2} wiggly line represents energy that is partly lost by convection and partly used when water in the sea turns to water vapour.



Exercises

22 Using the data on the Sankey diagram, calculate the albedo of the Earth.

8.10 Global warming

Assessment statements

8.6.1 Describe some possible models of global warming.
8.6.2 State what is meant by the enhanced greenhouse effect.
8.6.3 Identify the increased combustion of fossil fuels as the likely major cause of the enhanced greenhouse effect.
8.6.4 Describe the evidence that links global warming to increased levels of greenhouse gases.
8.6.5 Outline some of the mechanisms that may increase the rate of global warming.

We can see from our model of the greenhouse effect that the temperature of the Earth depends on several factors. For example, if the amount of radiation coming from the Sun were to increase then there would be more energy reaching the Earth. This would cause the temperature of the Earth to increase until equilibrium is restored; this is called *global warming*.

Models for global warming

The factors affecting the temperature of the planet are very complex and interrelated. Physicists use their knowledge to make mathematical models so that they can predict the outcome of changing variables. However, this problem is rather more difficult than an ideal gas or any other system considered so far in this course, and the equations are equally difficult to solve. To solve these equations, physicists make computer models and program computers to do the millions of calculations required. To show how these are built up, we will consider an analogous situation.

Sand analogy

It is difficult to imagine what is happening to all the energy flowing in and out of the Earth. To make this easier to visualize, we can consider an analogous situation such as loading sand onto a truck.

Imagine you are filling a truck with sand. As you put sand onto the truck someone else takes off a fixed percentage of the complete load. As the amount on the truck increases, they take off more until they take off as much as you put on – an equilibrium is reached.

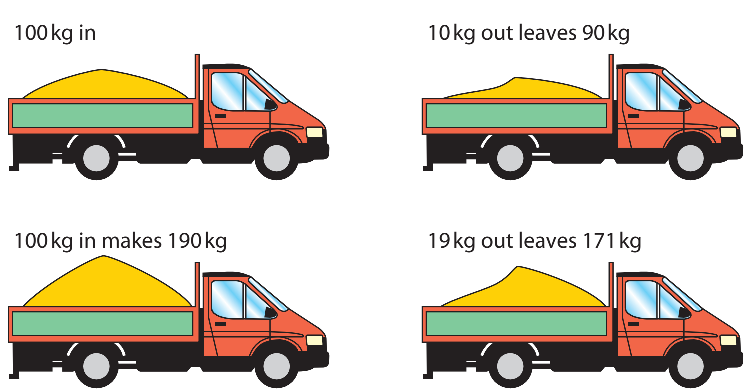


Figure 8.40 In the sand analogy, if 100 kg are put in each minute and 10% is taken out each minute then the outcome would be as shown.

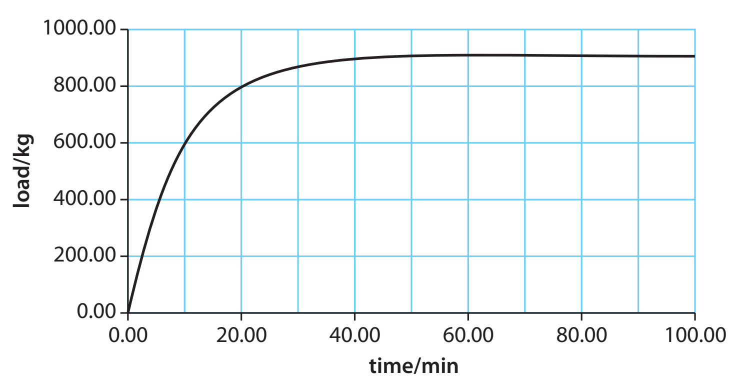
To find out how the load varies until equilibrium is reached we have to do a lot of calculations, or we can use a spreadsheet to do it for us (See Figure 8.41). Once the data has been entered into a spreadsheet it is a simple matter of copying the formula down in as many rows as you want. If you do this, then make sure you copy row 3 not row 2, since row 2 only contains zeros. If this is copied down to 100 minutes you will see that equilibrium has been reached when the load contains 900 kg, as every time another 100 kg is added, 100 kg will be taken out. Once you have made this model it is very easy to change the variables to see what would happen. You can try adding more sand by changing column D or taking more out by changing the factor 0.1 in column E.

Figure 8.41 shows how the sand analogy would be put into a spreadsheet and how the results can be displayed as a graph of load against time.

The amount taken out is 0.1 × 100
= 0.1 *(D3 + F2)

The amount left is what was there
+ what was added
− what was taken out
= 0 + 100 − 10 = 90 = F2 + D3 − E3

	C	D	E	F
1	Time (mins)	In (kg)	Out (kg)	Load (kg)
2	0	0.00	0.00	0.00
3	2	100.00	10.00	90.00
4	2	100.00	19.00	171.00
5	3	100.00	27.10	243.90
6	4	100.00	34.39	309.51



Global warming is not as simple as the sand model but it's the same principle.

Modelling global warming

The energy flow for the Earth and atmosphere is rather more complicated than the sand analogy. One complication is the greenhouse effect, which would be like putting some of the sand back into the truck. Computer simulations that model the climate of the Earth must take all factors into consideration, and that takes a lot more computer power than even the fastest home computer. One concerned group, 'Climate prediction', have been doing calculations on thousands of private computers in homes and schools all around the world to gain the power needed to run their computer model.

W

To see a simple spreadsheet model for the Earth (without the greenhouse effect), visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 8.9. Here you will also find a lot of information about climate change and details of the other simulations they are running.

Exercises

23 Try making the spreadsheet model as above. See what happens if you change the variables.

Causes of global warming

By analysing the energy flow diagram in Figure 8.38 we can see that there are several ways that the temperature of the Earth could increase.

The radiation from the Sun

The radiation from the Sun is not constant. If the amount of radiation incident on the Earth increases, then its temperature would increase. There are several factors that affect this:

Solar flares/sunspots

Sunspots are black spots on the surface of the Sun that can be seen if you look at the Sun through a sufficiently dark filter. The spots are cool areas. However, when there are a lot of sunspots, the Sun emits more energy due to the increased temperature of the gas surrounding the spot. The number of sunspots varies on an 11-year cycle.

The increase in temperature as a result of many sunspots together leads to solar flares, jets of gas flying out from the Sun's surface like huge flames.

Earth's orbit

The Earth's orbit is not circular but elliptical; this means that the distance between the Earth and the Sun is not constant. There are also other variations due to the change of angle of the Earth's axis in relation to the Sun. These are called Milankovitch cycles.

Enhanced greenhouse effect

If the amount of greenhouse gases in the atmosphere is increased, then the amount of radiation absorbed by the atmosphere increases. The amount of energy leaving the Earth is reduced and the temperature of the Earth would rise until equilibrium was restored.

The greenhouse gases include water, carbon dioxide, methane and nitrous oxide. Water is the biggest contributor to the greenhouse effect, with carbon dioxide second. Methane has a much bigger effect, but there isn't so much of it in the atmosphere.

Greenhouse gases
There are many gases that contribute to the greenhouse effect and the complete picture of their combined contribution is very complicated and out of the scope of this book. To simplify matters we will only deal with CO_2 since this is the major contributor.

Ice cores

For hundreds of thousands of years the ice of Antarctica has been growing. Each year a new layer is added on top of the old, so that the ice that is there now is made up of thousands of layers, each layer representing a year's growth (like the rings of a tree). By drilling into the ice with a hollow drill it is possible to extract samples (ice cores) that were laid down thousands of years ago. From the concentration of the different isotopes of hydrogen in the water it is possible to determine the temperature of the layers: more heavy isotopes means the temperature was colder. These layers of ice also contain bubbles of air that have been trapped since the ice was laid down; from these bubbles we can find how the composition of the atmosphere has changed since the ice was formed. If we compare the temperature of the Earth with the concentration of CO_2 we get an interesting result, as shown in Figure 8.42. Comparing these two graphs we can see that when the concentration of CO_2 increases, so does the temperature.

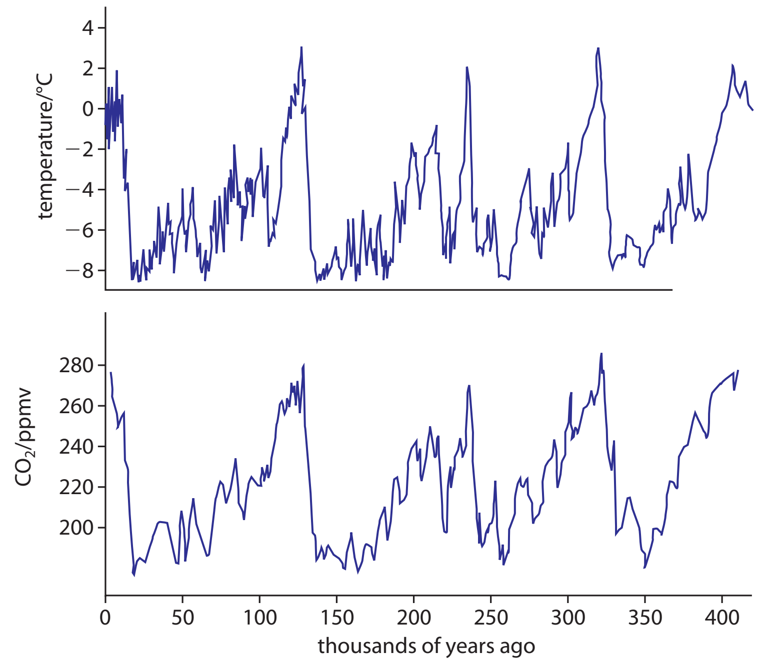


Figure 8.42 Data from the Vostok Antarctic ice core can be explained using the greenhouse gas model, however this piece of evidence alone is not enough to say that temperature depends on CO_2 concentration.

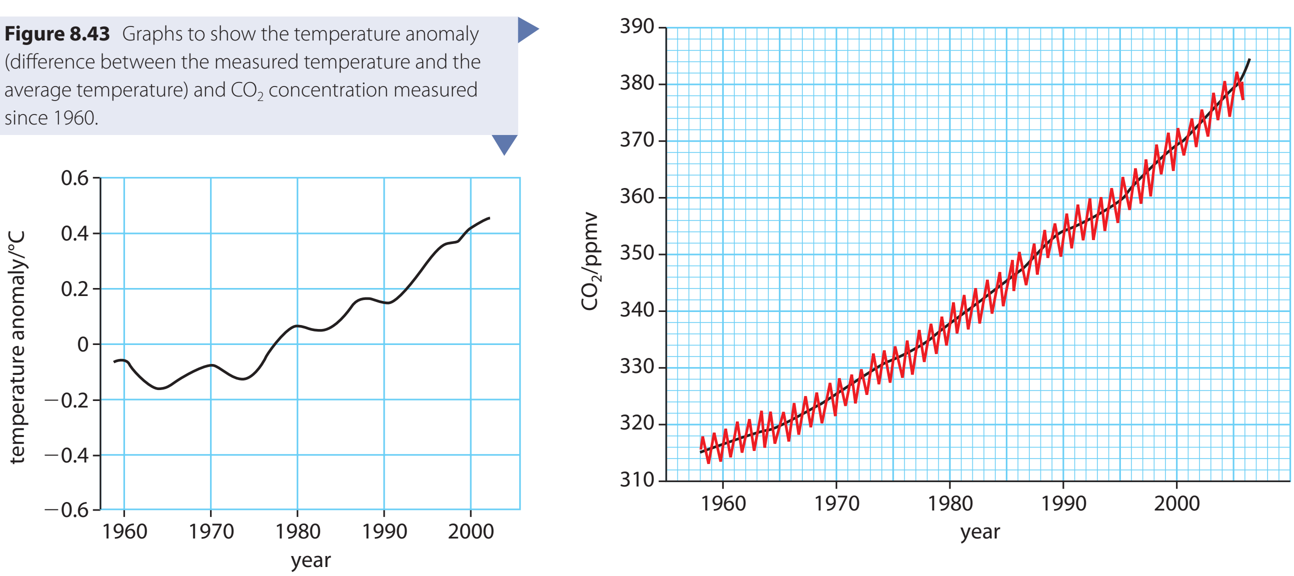
If the results from the ice core were a class practical, then you would devise an experiment to test the hypothesis that, when the concentration of CO_2 increases so does the temperature (making sure that the variables were controlled). It's of course not possible to experiment with the amount of CO_2 in the atmosphere so computer simulations may be used instead.

The temperature scale is the difference in temperature between the temperature calculated from the hydrogen isotopes in the ice core and the average temperature now.
To find data from the ice cores, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on the Weblink 8.10.

More recent data

From measurements of the average temperature of the Earth and the amount of CO_2 in the atmosphere, it can be seen that there could be a relationship between the two. The graphs in Figure 8.43 show both have risen in the past 50 years. Notice there is regular yearly variation in CO_2, the maximum coming after the northern hemisphere winter; this is because during the summer, plants absorb CO_2 from the atmosphere.

Figure 8.43 Graphs to show the temperature anomaly (difference between the measured temperature and the average temperature) and CO_2 concentration measured since 1960.



What causes the change in CO_2?

It is widely believed that the increase in CO_2 is due to human activity, mainly because of the burning of fossil fuels which produces the gas, and deforestation which removes plants that would normally absorb the gas. This hypothesis is supported by the increased use of fossil fuels during the past 50 years (see Figure 8.44).

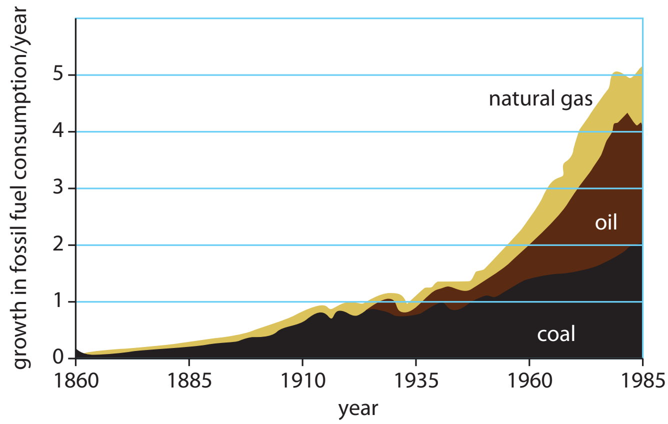


Figure 8.44 The exponential increase in fossil fuel consumption has contributed to the increase in CO_2 in the atmosphere.

8.11 What might happen and what can be done?

Assessment statements

8.6.6 Define *coefficient of volume expansion*.

8.6.7 State one possible effect of the enhanced greenhouse effect

8.6.8 Outline possible reasons for a predicted rise in mean sea-level.

8.6.9 Identify climate change as an outcome of the enhanced greenhouse effect.

8.6.10 Solve problems related to the enhanced greenhouse effect.

8.6.11 Identify some possible solutions to reduce the enhanced greenhouse effect.

8.6.12 Discuss international efforts to reduce the enhanced greenhouse effect.

The Steigletscher glacier in Switzerland photographed in 1994 (top) and 2006 (bottom). Glaciers are shrinking like this all round the world.

To find out what would happen if the average temperature of the Earth were to increase, we can look at what has happened during the past 50 years, especially global warming in recent years, or we can use computer simulations.

Rise in sea level

As the temperature of a liquid increases, it expands. The relationship between the increase in volume (ΔV) and the temperature change (ΔT) is given by the formula

$$\Delta V = \beta V_0 \Delta T$$

where β = coefficient of volume expansion

V_0 = the original volume

If this is applied to water, then we can conclude that if the average temperature of the oceans increases then they will expand. This has already been happening; over the past 100 years sea level has risen by 20 cm. Trying to predict what will happen as the sea temperature increases is complicated by the anomalous expansion of water. Unlike a lot of other liquids, water does not expand uniformly, in fact from 0 °C to 4 °C water actually contracts and then from 4 °C upwards it expands. Trying to calculate what happens as different bodies of water expand and contract is very difficult, but most models predict some rise in sea level.

Positive feedback
Melting ice caps have dual impact. We have discussed the way that the amount of radiation reflected off the Earth is an important factor in determining the Earth's temperature. The ice caps are white and therefore reflect a high amount of radiation (their albedo is high). If they melt, then the amount of reflected radiation would be reduced, causing the temperature to rise further; this is called *positive feedback*.

Another factor that could cause the sea level to rise is the melting of the ice caps. This is the ice that covers the land masses of Antarctica, Greenland and the glaciers in mountainous regions. It doesn't include the Arctic since this ice is floating. Floating ice displaces its own mass of water so when it melts it makes no difference. If the ice caps melt and the water runs into the sea, then it could make the sea level rise so much that some countries could disappear under water.

Change in the weather

The other obvious consequence of the enhanced greenhouse effect is a change in the weather. What would happen is, again, difficult to predict but most models agree that countries near the equator will get hotter and countries in the northern hemisphere will get wetter.

Solutions

To reduce the enhanced greenhouse effect, the levels of greenhouse gases must be reduced, or at the very least, the rate at which they are increasing must be slowed down. There are several ways that this can be achieved:

1 **Greater efficiency of power production**
In recent years the efficiency of power plants has been increasing significantly. According to the second law of thermodynamics, they can never be 100% efficient but some of the older less efficient ones could be replaced. This would mean that to produce the same amount of power would require less fuel, resulting in reduced CO_2 emission.

2 **Replacing the use of coal and oil with natural gas**
Gas-fired power stations are more efficient than oil and gas and produce less CO_2.

3 **Use of combined heating and power systems (CHP)**
Using the excess heat from the power station to heat homes would result in a more efficient use of fuel.

4 **Increased use of renewable energy sources and nuclear power**
Replacing fossil fuel burning power stations with alternative forms such as wave power, solar power and wind power would reduce CO_2 emissions.

5 **Use of hybrid vehicles**
A large amount of the oil used today is used for transport, and even without global warming, there will be a problem when the oil runs out. Cars that run on electricity or a combination of electricity and petrol (hybrid) are already in production. Aeroplanes will also have to use a different fuel.

6 **Carbon dioxide capture and storage**
A different way of reducing greenhouse gases is to remove CO_2 from the waste gases of power stations and store it underground.

An international problem

Global warming is an international problem, and if any solution is going to work then it must be a joint international solution. Before working on the solution the international community had to agree on pinpointing the problem and it was to this end that the Intergovernmental Panel on Climate Change (IPCC) was formed.

IPCC

In 1988 the World Meteorological Organisation (WMO) and the United Nations Environmental Programme (UNEP) established the IPCC, the panel of which was open to all members of the UN and WMO. Its role was not to carry out research but to assess all the available information relating to human induced climate change.

An excerpt from the first report of Working Group I in 1990 states:
"The experts concluded that they are certain that emissions from human activities are substantially increasing the atmospheric concentrations of greenhouse gases and that this will enhance the greenhouse effect and result in an additional warming of the Earth's surface."

Kyoto Protocol

In 1997 the Kyoto Protocol was open for signature: countries ratifying this treaty committed to reduce their greenhouse gas emissions by given percentages. By January 2009, 181 countries had signed and ratified.

Asia-Pacific Partnership on Clean Development and Climate (APPCDC)

This is a non-treaty agreement between six countries that account for 50% of the greenhouse emissions. The countries involved agreed to cooperate on the development and transfer of technology with the aim of reducing greenhouse emissions.

W To access the websites of the IPCC, the UN and the Asia-Pacific Partnership, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblinks 8.11, 8.12 and 8.13.

Practice questions

1 This question is about energy sources.

(a) Fossil fuels are being produced continuously on Earth and yet they are classed as being non-renewable. Outline why fossil fuels are classed as non-renewable. (2)

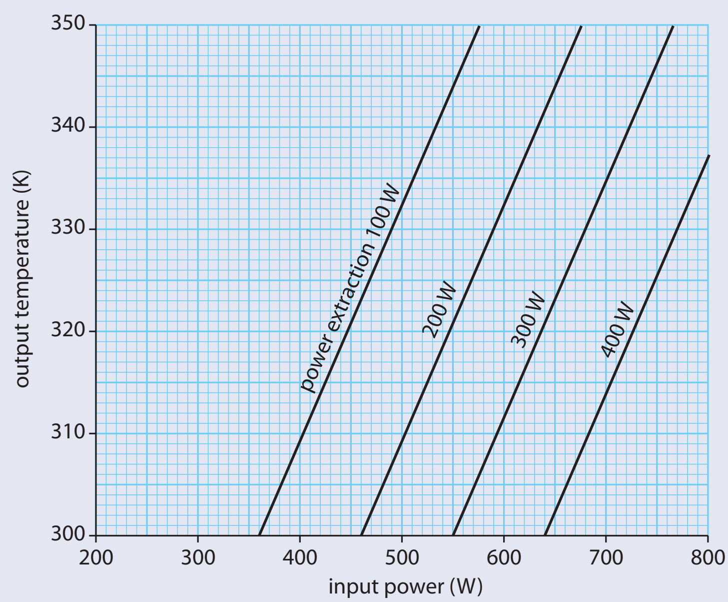
(b) Some energy consultants suggest that the solution to the problem of carbon dioxide pollution is to use nuclear energy for the generation of electrical energy. Identify **two** disadvantages of the use of nuclear fission when compared to the burning of fossil fuels for the generation of electrical energy. (2)

(Total 4 marks)

2 This question is about solar energy.

(a) By reference to energy transformations, distinguish between a solar panel and a solar cell. (2)

Some students carry out an investigation on a solar panel. They measure the output temperature of the water for different solar input powers and for different rates of extraction of thermal energy. The results are shown below.



(b) Use the data from the graph to answer the following.

(i) The solar panel is to provide water at 340 K whilst extracting energy at a rate of 300 W when the intensity of the sunlight incident normally on the panel is 800 W m^{-2}. Calculate the effective surface area of the panel that is required. (2)

(ii) Deduce the overall efficiency of the panel for an input power of 500 W at an output temperature of 320 K. (3)

(Total 7 marks)

3 This question is about the production of electrical energy.

(a) Outline the principal energy transfers involved in the production of electrical energy from thermal energy in a coal fired power station. (2)

(b) State and explain whether the energy sources used in the following power stations are renewable **or** non-renewable.

(i) Coal fired (1)

(ii) Nuclear (1)

(c) The core of some nuclear reactors contains a moderator and control rods. Explain the function of these components.

(i) The moderator (2)

(ii) The control rods (2)

(d) Discuss **one** advantage of a nuclear power station as opposed to a coal-fired power station. (2)

(Total 10 marks)

4 This question is about wind energy.

It is required to design wind turbines for a wind farm for which the following information is available.

Total required annual electrical energy output from the wind farm $= 120$ TJ

Maximum number of turbines for which there is space on the farm $= 20$

Average annual wind speed at the site $= 9.0 \text{ m s}^{-1}$

(a) Deduce that the average power output required from one turbine is 0.19 MW. (3)

(b) Estimate the blade radius of the wind turbine that will give a power output of 0.19 MW. (Density of air $= 1.2 \text{ kg m}^{-3}$) (3)

(c) State **one** reason why your answer to **(b)** is only an estimate. (1)

(d) Discuss briefly **one** disadvantage of generating power from wind energy. (2)

(Total 9 marks)

5 This question is about the production of nuclear energy and its transfer to electrical energy.

(a) When a neutron "collides" with a nucleus of uranium-235 (U) the following reaction can occur.

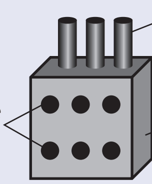
$$^{235}_{92}\text{U} + {}^{1}_{0}\text{n} \rightarrow {}^{144}_{56}\text{Ba} + {}^{90}_{36}\text{Kr} + 2\,{}^{1}_{0}\text{n}$$

(i) State the name given to this type of nuclear reaction. (1)

(ii) Energy is liberated in this reaction. In what form does this energy appear? (1)

(b) Describe how the neutrons produced in this reaction may initiate a chain reaction. (1)

The purpose of a nuclear power station is to produce electrical energy from nuclear energy. The diagram below is a schematic representation of the principle components of a nuclear reactor "pile" used in a certain type of nuclear power station.



The function of the moderator is to slow down neutrons produced in a reaction such as that described in part **(a)** above.

(c) **(i)** Explain why it is necessary to slow down the neutrons. (3)

(ii) Explain the function of the control rods. (2)

(d) Describe briefly how the energy produced by the nuclear reactions is extracted from the reactor pile and then transferred to electrical energy. (4)

(Total 12 marks)

6 This question is about nuclear power and thermodynamics.

(a) A fission reaction taking place in the core of a nuclear power reactor is

$${}^{1}_{0}\text{n} + {}^{235}_{92}\text{U} \rightarrow {}^{144}_{56}\text{Ba} + {}^{89}_{36}\text{Kr} + 3\,{}^{1}_{0}\text{n}.$$

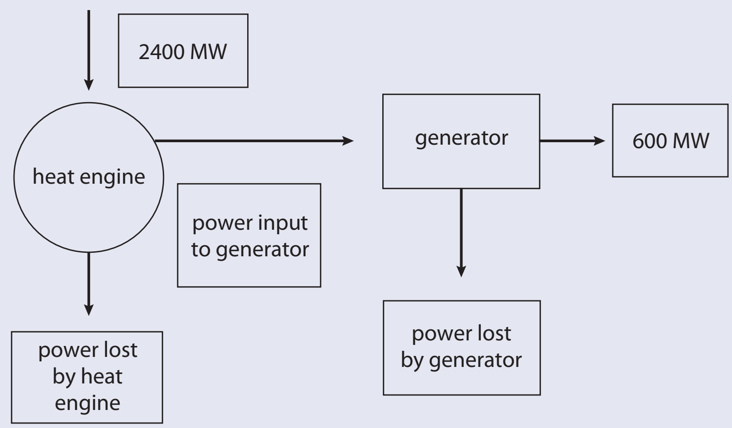
(i) State **one** form in which energy is released in this reaction. (1)

(ii) Explain why, for fission reactions to be maintained, the mass of the uranium fuel must be above a certain minimum amount. (2)

(iii) The neutrons produced in the fission reaction are fast moving. In order for a neutron to fission U-235 the neutron must be slow moving. Name the part of the nuclear reactor in which neutrons are slowed down. (1)

(iv) In a particular reactor approximately 8.0×10^{19} fissions per second take place. Deduce the mass of U-235 that undergoes fission per year. (3)

(b) The thermal power from the reactor is 2400 MW and this is used to drive (operate) a heat engine. The mechanical power output of the heat engine is used to drive a generator. The generator is 75 % efficient and produces 600 MW of electrical power. This is represented by the energy flow diagram below.



(i) Calculate the power input to the generator. (1)

(ii) Calculate the power lost from the generator. (1)

(iii) Calculate the power lost by the heat engine. (1)

(iv) State the name of the law of physics which prohibits **all** of the 2400 MW of input thermal power from being converted into mechanical power. (1)

(vi) The heat engine operates in a Carnot cycle with a low temperature reservoir of 300K. Calculate the temperature of the hot reservoir. (2)

(Total 14 marks)

7 This question is about wind power.

(a) A wind turbine produces 15 kW of electric power at a wind speed v.

(i) Assuming a constant efficiency for the wind turbine, determine the power output of the turbine for a wind speed of $2v$. (2)

(ii) Suggest **two** reasons why all the kinetic energy of the incident wind cannot be converted into mechanical energy in the turbine. (2)

(b) State and explain **one** advantage of using wind power to generate electrical energy as compared to using fossil fuels. (2)

(Total 6 marks)

9 Digital technology

9.1 Analogue and digital signals

Assessment statements

14.1.1 Solve problems involving the conversion between binary numbers and decimal numbers.

14.1.2 Describe different means of storage of information in both analogue and digital forms.

14.1.3 Explain how interference of light is used to recover information stored on a CD.

14.1.4 Calculate an appropriate depth for a pit from the wavelength of the laser light.

14.1.5 Solve problems on CDs and DVDs related to data storage capacity.

14.1.6 Discuss the advantage of the storage of information in digital rather than analogue form.

14.1.7 Discuss the implications for society of ever-increasing capability of data storage.

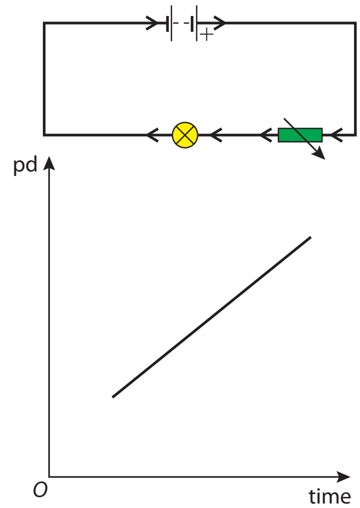


Figure 9.1 An analogue signal.

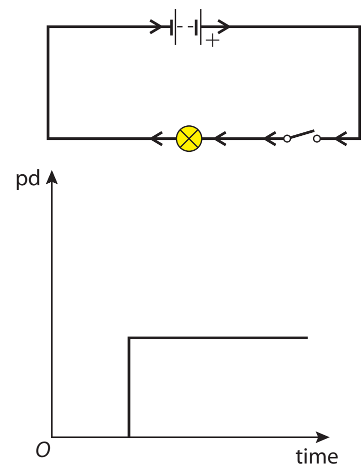


Figure 9.2 A digital signal.

In this topic we will be learning about the physical principle behind the operation of many of today's digital devices, the digital camera, data storage on CD and DVD and the mobile phone system. Before starting this section you will need to have covered the basic electricity and waves sections in the core chapters.

Electrical signals

When dealing with electronic appliances we will often refer to a *signal*. A signal is the transfer of information from one place to another. For example, if you wave your hand to a friend, that is a signal from one person to another. An electrical signal is when information is sent via a changing electric field; a simple example would be if the switch in Figure 9.2 is opened and closed so that a changing electric field is sent to a light bulb.

Analogue signal

An analogue signal is a continuously varying signal. For example, in the circuit shown in Figure 9.1, the current changes from a minimum to a maximum as the resistance is decreased, causing the bulb get progressively brighter. The change in potential difference (p.d.) is an analogue signal.

Digital signal

A digital signal is not a continuously varying signal but one that changes from one value to another. If we consider the circuit in Figure 9.2 the bulb is off but if the switch is closed the bulb will suddenly glow bright. This step from low potential to high is a digital signal.

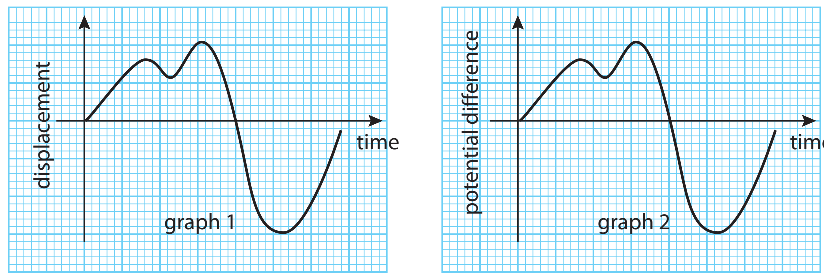
Transmitting sound

A more complicated signal could be sent using a microphone. The change in pressure caused when you speak makes a part inside the microphone vibrate, resulting in a changing p.d. The p.d. across the microphone varies in exactly the same way as the pressure of the sound wave. A graph of p.d. against time would therefore have the same shape as a graph of displacement against time

This signal can then be fed to an amplifier which increases the p.d.. If the p.d. is now connected across a loudspeaker, the loudspeaker will reproduce the original sound but louder.

The electrical signal between microphone and speaker is an analogue signal because it is a continuously varying signal.

Figure 9.3 Graph1 shows the displacement of the microphone and Graph 2 the p.d. produced.



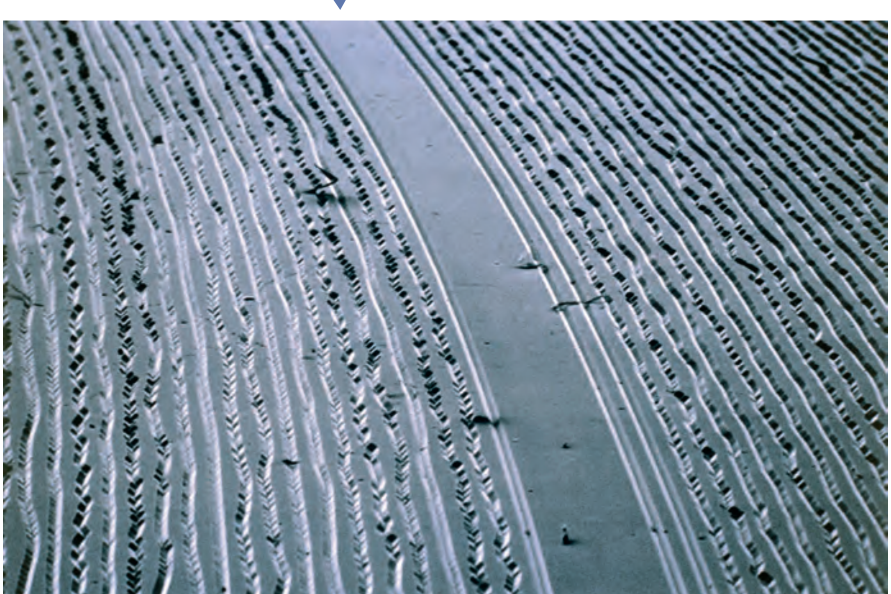
Recording an analogue signal

A sound can be stored on a vinyl record or an audio tape. These are both examples of analogue recording.

Vinyl record (LP)

The grooves in a vinyl record have the same shape as a graph of the original sound.

A record is a disc made of plastic (vinyl) that has a thin groove cut in it by a needle that moves from side to side at the same frequency as the sound. If you look at the groove under a microscope it looks like a graph of the sound wave. The record is played on a turntable with a needle resting in the groove. As the record rotates the needle is made to vibrate by the wavy groove at the same frequency as the original sound. The needle is connected to a device that turns this vibration back into an electrical signal, which can then be amplified and transmitted to a loudspeaker. The groove in the record spirals inwards so the needle moves through the groove faster when it is playing music on the edge of the LP. This means that the wavy groove is more squashed in the centre of the record than at the outer edge.



Audio tape

Audio tape is a magnetic tape wound between two spools in a cassette. It is reeled past an electromagnet which varies its magnetic field at the same rate as the signal. The result is that the tape gains a magnetic field that varies continuously at the same rate as the signal.

We say that the audio tape stores an analogue signal – in other words, it is continuously varying. In fact, it cannot be continuously varying since the tape is not continuous; it is made of molecules.

Digital devices

Many of today's electronic devices are said to be digital, e.g. computers, MP3 players and some telephones. These devices contain a microprocessor that can perform many functions. For example, an MP3 player not only plays music but tells you the name of the track and the artist. A DVD player can give subtitles in many different languages and a computer can do amazing things to a photograph. The microprocessor contains millions of tiny switches that can be either on or off. When they are on they give out a signal and when off there's no output. They are able to process digital signals.

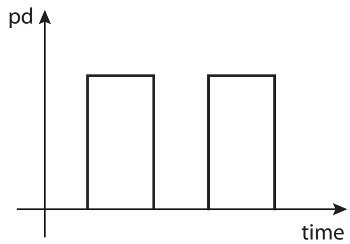


Figure 9.4 A digital signal can be represented mathematically by a series of 1s and 0s. The high part of the signal is a 1 and the low a 0.

Digital electronics

To understand the significance of using digital signals we will have a look at some simple digital electronics.

Logic gates

Digital circuits, for example the microprocessor in a computer, are made up of logic gates; these give out different signals depending on the input. There are two main types of gate: the AND gate and the OR gate (see Figures 9.5 and 9.6)

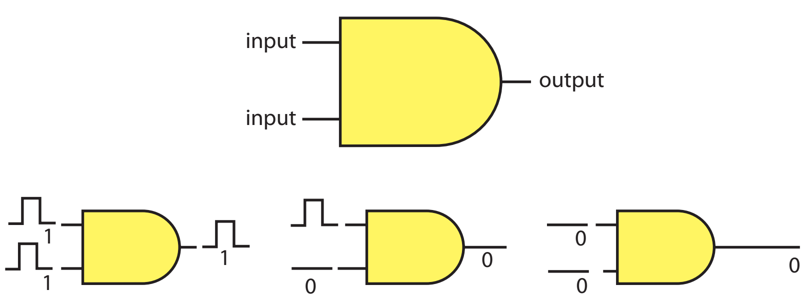


Figure 9.5 The AND gate symbol. An AND gate has two inputs and one output. It only gives an output if there is a signal on each input. It could be used to switch on a light, for example, only if two switches are closed.

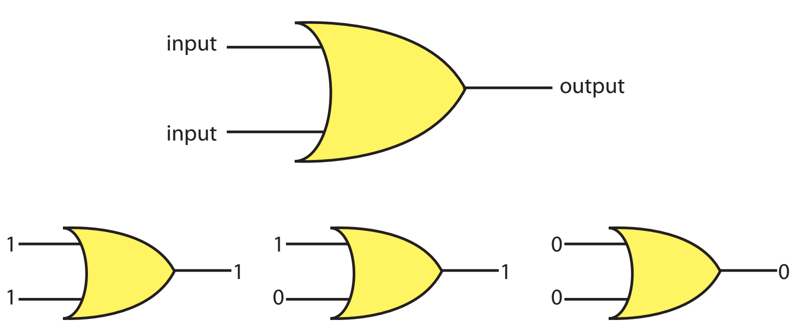


Figure 9.6 The OR gate symbol. This gate gives an output if there is a signal on one input or the other. It could be used to switch on a light when either one or both switches are closed.

By using combinations of different types of logic gate, many processes can be performed that are much more complicated than simply switching lights on and off, for example making an MP3 player play the right track when you press certain buttons. Logic gates are simple devices and they only work with simple digital signals that are made of 1s and 0s. This may not seem very useful until you realise that all numbers can be made out of a series of 1s and 0s. This is called the *binary system*.

Binary numbers

Our normal system of numbers is called the *decimal system*. In this system we can use 10 symbols (0–9) to represent any number. The way it works is that we group units, tens, hundreds and so on. So 365 is 3 hundreds, 6 tens and 5 units. Units, tens and hundreds are all powers of 10: 10^0, 10^1, 10^2 etc. We can represent this in Table 1.

Table 1 The decimal system of numbers.

Power of 10	3	2	1	0
	1000	100	10	1
	thousands	hundreds	tens	units

The base of our number system doesn't have to be 10. We can use any number. If we use 2, this is called the binary system, as represented in Table 2.

Table 2 The binary system of numbers.

Power of 2	3	2	1	0
	8	4	2	1
	eights	fours	twos	ones

In this system the number 7 would be one 4, one 2 and one 1. See Table 3 for the numbers 1–10 in binary. With 4 digits we can count from 0 to 15, with 8 digits we can count to 255.

Table 3

decimal	binary
1	00000001
2	00000010
3	00000011
4	00000100
5	00000101
6	00000110
7	00000111
8	00001000
9	00001001
10	00001010

Bit (b)
Each 1 or 0 is called a binary digit or 'bit'.

Byte (B)
A byte is 8 binary digits. 8 digits can represent any number from 0 to 255.

Exercises

1 Convert the following numbers into 8-digit binary:
(a) 7 **(b)** 14 **(c)** 16 **(d)** 67 **(e)** 125

2 Convert the following 6-bit binary numbers to base 10 numbers:
(a) 000111 **(b)** 100100 **(c)** 110011 **(d)** 111111

ASCII code

It is not at all obvious that a computer uses 1s and 0s to perform its functions. After all, we don't communicate with a computer in this way, we type letters on a keyboard. What the computer does is to convert each letter into an 8-bit binary code. This is called the ASCII code, which stands for American Standard Code for Information Interchange. A small part of it is shown in Table 4 where each letter is 8 bits, or 1byte, of information. You can test this by typing the word *computer* into a notepad document and saving it on your computer. If you look at the properties of this file you will see that it has a file size of 8 bytes with one byte for each letter. If you do the same exercise in MS Word you will find the file size is much bigger because of all the formatting.

ASCII code	
A	0100 0001
B	0100 0010
C	0100 0011
D	0100 0100
E	0100 0101

Table 4 The ASCII code uses 8 bits to represent each character.

Why don't we communicate in binary language?

Chinese has tens of thousands of characters so each one cannot be represented by one byte of code. Two bytes are used instead, giving about 65 000 possible combinations.

Exercises

3 Write the word BAD in ASCII code.

4 If you typed the letter A once every second into a notepad document, how long would it take you to fill a 200 GB hard drive?

Analogue to digital conversion

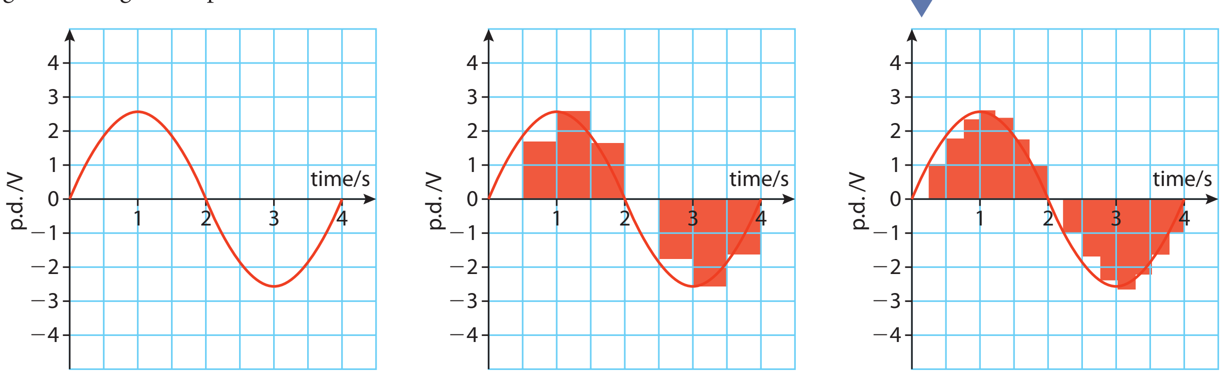
We have seen that a digital signal is more versatile than an analogue signal. It would, therefore, be a great advantage if we could change analogue signals, such as music, into digital signals. This can be done using an analogue to digital converter or ADC.

Sampling an analogue signal

The first stage in converting an analogue signal to a digital one is to measure the voltage at regular intervals to create a pulse amplitude modulated (PAM) signal. This is called *sampling*. Consider the signal in Figure 9.7. The first sample is rough but if the sampling rate is increased then the digital signal begins to look like the original analogue sample.

Sampling rate
The sampling rate is the rate at which a signal is measured. If the sampling rate is higher, the digital signal will be closer to the original sound.

Figure 9.7 An analogue signal is sampled at twice a second and then four times a second.



Nyquist frequency

The Nyquist theory states that to record a certain frequency the sampling rate must be at least twice the frequency. This is so that at least one point on each peak and trough is recorded per cycle. Sound is often sampled at a rate of 40 000 Hz so the highest frequency that can be recorded is 20 000 Hz, which is the highest frequency that humans can hear.

Exercises

5 Figure 9.8 shows a graph of the varying p.d. from a microphone. Convert this signal to a PAM signal by measuring the p.d. at a sampling rate of 200 Hz. Round off all of the p.d.s to the nearest whole volt. Draw a graph showing the digital signal.

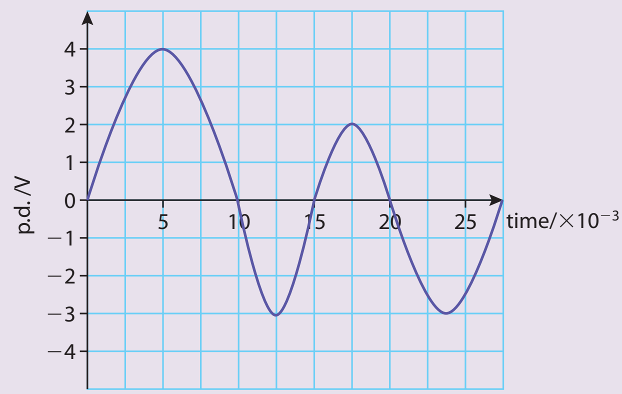


Figure 9.8

Does this sampling rate give a true representation of the signal?

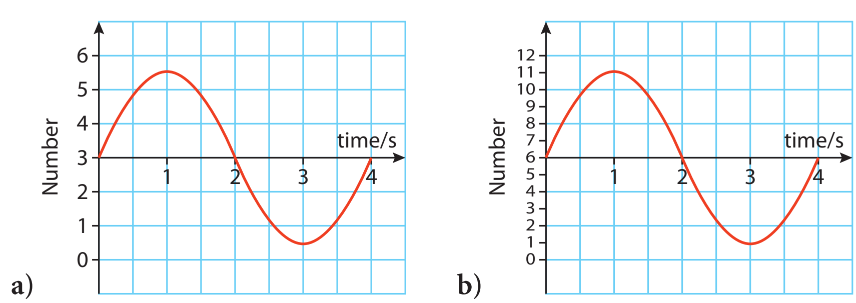
6 Repeat with a sampling rate of 400 Hz.

7 To measure the variation of a 50 Hz AC signal using a digital device, what is the minimum sampling rate you should use?

Changing to binary

PAM signals cannot be processed by a microprocessor until they are converted to the binary system. To change a PAM signal to binary we must split the p.d. into different levels which can then be converted to a binary number. The number of levels depends upon how many bits we are going to use to represent the number.

Figure 9.9 The p.d. of the example signal is split into **a)** 6 levels for a 3-bit digital conversion and **b)** 12 levels for a 4-bit digital conversion.



Number of bits

If we consider the signal in Figure 9.7 we see that it goes from about −3 V to +3 V. This range can most easily be represented by the numbers 0–6, each number representing a different voltage, as in Figure 9.9 . If we convert these numbers to binary they can be represented by 3 bits. If we sample this signal at a sampling rate of 1 Hz then we reach the values given in Table 5.

For 3-bit digital conversion		
Time	Number	Binary
0	3	011
1	6	110
2	3	011
3	0	000
4	3	011

Table 5 Binary values for an analogue signal for 3-bit digital conversion. The maximum number of levels for 3-bit conversion is 8.

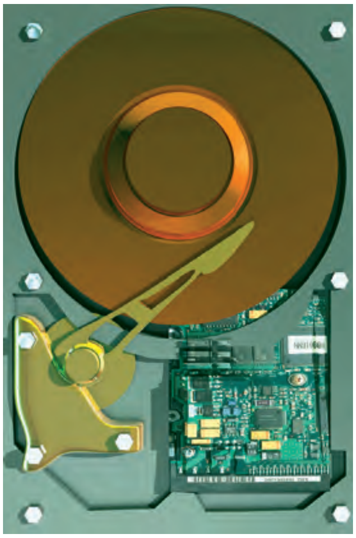
The complete signal for the 3-bit conversion is 011110011000011. This string of 1s and 0s can be turned back into the original signal as long as we know how many bits were being used in the original conversion.

In the process of converting the p.d. to a number we have had to round off the values, for example at 1 second the number is between 5 and 6 when using levels 1–6 for 3 bits. It would be better if we had more levels but this would require more bits. With 4 bits we can represent numbers from 0–15, and this would enable us to split the voltage range into 12 numbers, as in Figure 9.9. This would give the values in Table 6.

For 4-bit digital conversion		
Time	Number	Binary
0	6	0110
1	11	1011
2	6	0110
3	1	0001
4	6	0110

Table 6 Binary values for an analogue signal for 4-bit digital conversion. The maximum number of levels for 4-bit conversion is 16.

So, to record the signal accurately we need a high sampling rate and a large number of bits. A typical MP3 player samples 16 bits at a rate of 44.1 kHz. This means that the signal can be split into 65 536 levels. This sounds very impressive, but remember the original analogue signal was continuous, you could say it had an infinite number of levels.



A computer hard drive stores information in microscopic magnets that can either point in an up or down direction.

Exercises

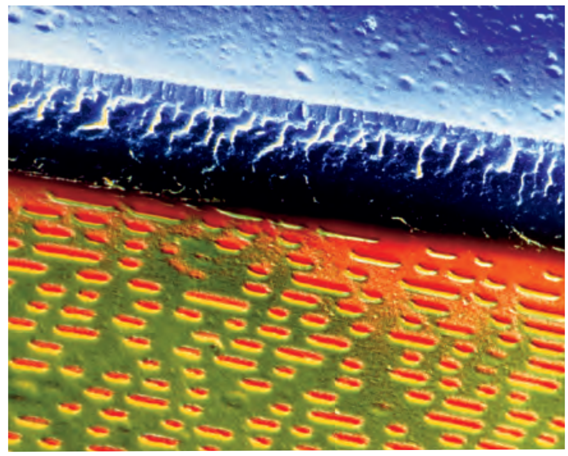
8 Convert the signal in the 12-level graph (Figure 9.9 b) to 4-bit binary with a sampling rate of 2 Hz. Present your data in a table similar to Table 6.

9 The following string of 1s and 0s is 3-bit binary data sampled at 2 Hz. Turn this data into a table giving time and a number representing the p.d. Use this data to draw the signal.
011100101100011000

Digital storage

A digital storage device, such as a computer hard drive or a CD, stores a series of 1s and 0s.

Compact disc (CD)



An electron microscope image of the pits that form the spiralling track on a CD.

Construction

A CD is a 12 cm plastic disc. Information is stored on it in the form of a spiralling track of small pits that are pressed into the plastic. To make the pits readable with a laser they are made shiny by spraying with a thin film of aluminium before being filled with an acrylic layer and a label. The result when viewed from the readable side is a shiny disc covered in microscopic bumps that spiral outwards from the centre in a 5 km long track.

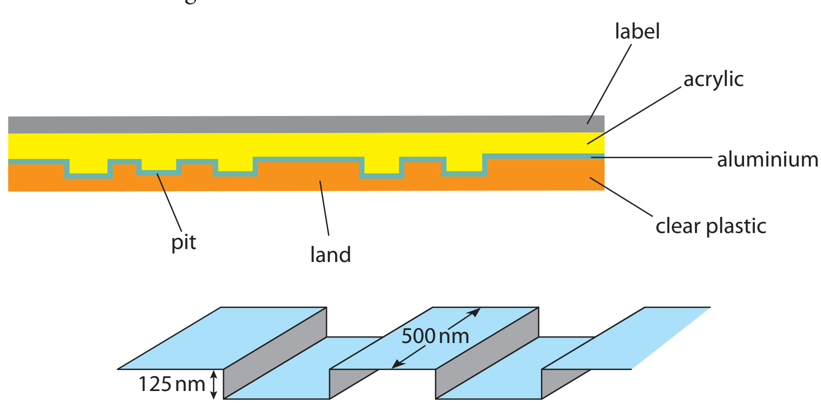


Figure 9.10 A diagram of a CD showing the dimensions and layers.

Reading the CD

The bumps are called pits because they are small holes in the plastic disc. When they are filled with aluminium, they form bumps. The lands are the parts in between.

The shiny bumps or pits are read using a laser. When it shines on the step at the edge of a pit, part of the beam reflects off the pit and part off the land. The height of the pit (125 nm) is $\frac{1}{4}$ of the wavelength of the laser light. (The laser light has a wavelength of 780 nm but when it passes into the plastic it slows down resulting in a decrease of the wavelength to 500 nm).

After reflection the two beams meet at a sensor located in the pickup. At this point the beam reflected off the bottom of the disc has travelled $\frac{1}{2}$ a wavelength further than the wave reflected off the top of the pit. The two waves are, therefore, out of phase and interfere destructively, causing the reflected light to 'blink'.

The sensor produces an emf that is proportional to the amount of light it receives. When the destructive interference occurs the emf goes down.

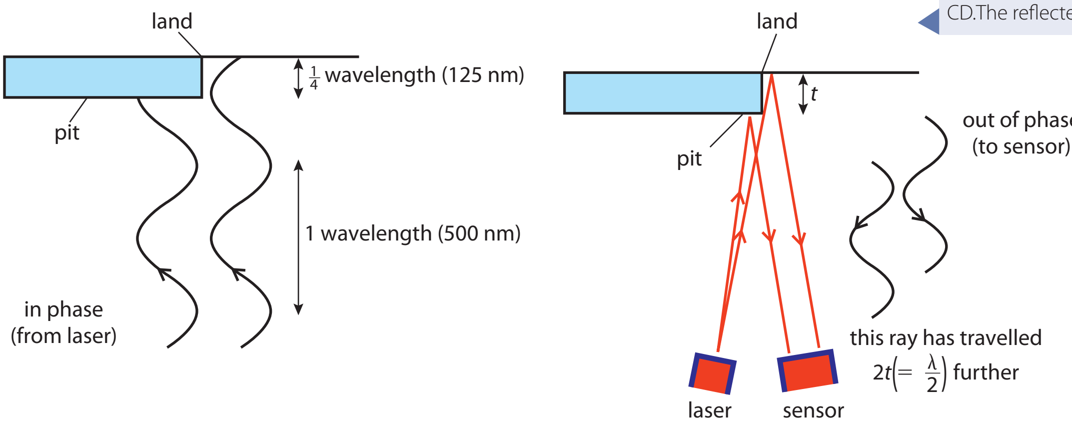


Figure 9.11 Two rays of laser light reflected off the edge of a pit on a CD.The reflected rays are out of phase.

Worked example

If the laser light used had a wavelength of 600 nm in plastic, what depth of pit would the CD have?

Solution

To achieve destructive interference, the path difference between light from the top of the pit and bottom of a land should be $\frac{1}{2}\lambda$. Since the light travels there and back the pit height must therefore be $\frac{1}{4}\lambda = \frac{600}{4} = 150$ nm.

Timing
It is important that the reader reads at just the right time. To synchronize the reader, the track is slightly off centre. The constant wobble this causes is used to keep the reader synchronized.

Tracking

When a CD is put into a CD player, it spins at about 700 rpm and the laser and sensor are moved slowly outwards, following the spiralling path of pits. The sensor samples at 44 100 times per second, which means that about every 23 μs a reading is taken. Whenever there is a change from a pit to a land, the laser blinks. This is represented by a 1 and if there is no blink then this is represented by a 0.

It is important that the laser reads the pits at a constant rate, otherwise it might miss some information or record more 'no blinks' than actually exist. If the CD rotated at a constant rate, the velocity would increase as the laser moved to the outer edge, so the CD slows down as the laser moves outwards. A CD can store 700 MB of information.

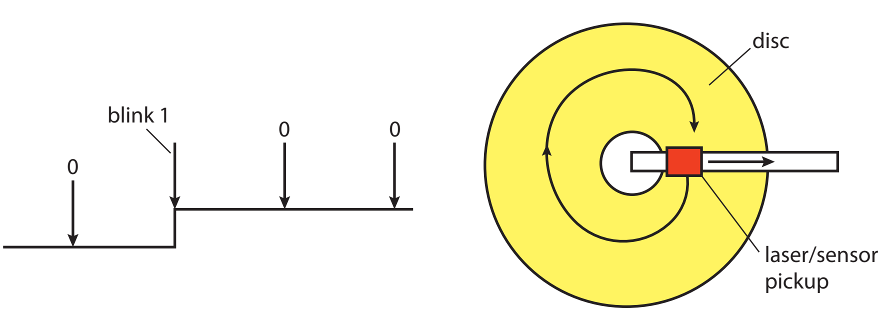


Figure 9.12 The laser tracks outwards. The reflected light is read at constant intervals:
blink – 1
no blink – 0

Worked example

If the CD rotates at a rate of 500 rpm when the pickup is 2 cm from the centre, how fast should it rotate when it is 4 cm from the centre?

Solution

We know from circular motion that the speed of a rotating body is given by:
$v = \omega r$ where $\omega = 2\pi f$

so $v = 2\pi fr$

If v is constant this means that $f \propto \frac{1}{r}$ so if the radius is doubled the frequency is halved to 250 rpm.

Digital video disc (DVD)

A DVD operates on the same physical principle but has a much higher storage capacity. This is achieved in several ways

Laser

The laser used to read a DVD has a wavelength of 640 nm, which means that smaller pits can be read.

Track

Each pit is about $\frac{1}{2}$ the size of the pits on a CD. In addition, the track spirals are twice as close, making the track length twice as long as that of a CD with twice as many pits on it.

Layers

It is possible to double the information on a disc by having two layers of pits. The top layer is coated with a semi-reflective coating enabling light to also pass through to read the bottom layer.

On a double sided double layer DVD it is possible to store 17 GB of data, a lot more than the 700 MB of a typical CD.

How many bytes?
1 page of writing 10 kB

1 metre book shelf of books 100 MB

Complete works of Beethoven 20 GB.

Exercises

10 The track on a CD is 5 km long and is made up of a series of pits and lands that are a minimum of 0.83 μm long.

(a) How many pits are there on the CD track (remember each pit is followed by a land)?

(b) Each short pit has two edges so represents 2 bits of data. How many bits are there on a CD?

(c) How many bytes of data are there on a CD?

11 A CD can store 74 minutes of stereo music.

(a) If 16 bits are recorded at 44 000 samples per second, how many bits are recorded in 74 minutes?

(b) The music is stereo so there are 2 channels. How many bits must the CD contain?

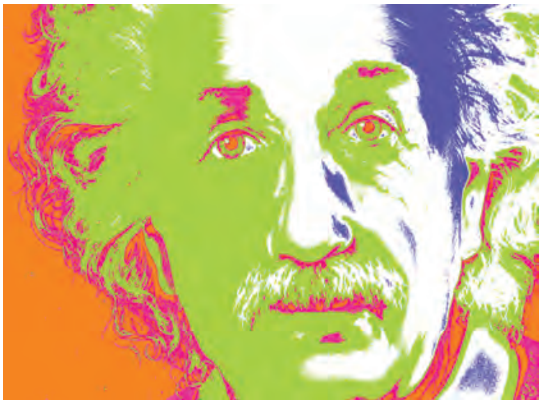
(c) How many bytes is that?

Comparing analogue and digital storage

There will always be a difference between an analogue signal and its digital equivalent. However, we must remember that when an analogue signal like a piece of music is stored on an analogue storage device such as an LP or audio tape, it will not be exactly the same as the original sound: the electronics needed to move the needle to cut the groove in an LP will make small changes to the signal. It is true that an analogue storage device does store more data than a digital device, but there is a point beyond which it is not possible to hear the difference between a digital recording and an analogue recording. However, once the information is stored, there are many advantages to storing digitally.

Processing data

A computer is a powerful tool that can process digital information. A clear advantage of storing data digitally is that you can use a computer to change it. For example, digital photographs can be enhanced or completely changed using Adobe Photoshop; a singer's voice can be edited to sing in tune; video clips can be joined and edited together.



A digitally manipulated photograph of Einstein.

All of these functions can be carried out on analogue data but the processes require specially made, expensive equipment. The reason that it is cheaper to process a digital signal is that the components do not have to be very precise, as the 1s and 0s can be represented by a high and a low potential, respectively. The high potential can vary but will still be high and therefore recognized as a 1, but changing an analogue signal by the same amount will make it unrecognizable.

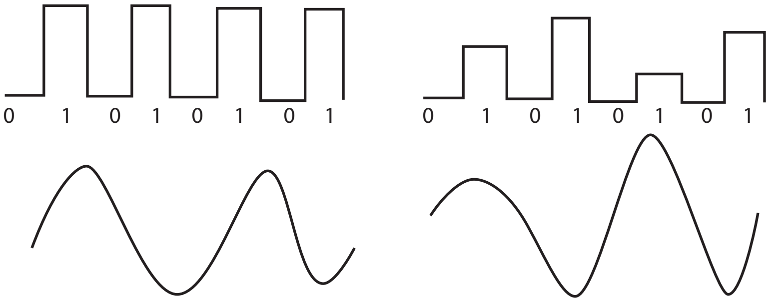


Figure 9.13 Even though the digital signal is changed the data is the same, unlike the analogue signal that is now very different.

Accessing data

One major advantage of an MP3 or CD player over either a tape recorder or record player is the possibility of jumping from track to track. The analogue data on an audio tape is stored in sequence, so to get to the last track on an album you have to wind the tape to the end. This is not a problem with a digital device where the data is stored like folders in a filing cabinet. This form of storage also makes it possible to combine different sorts of data, such as the language of the subtitles on a DVD.

Corruption of data

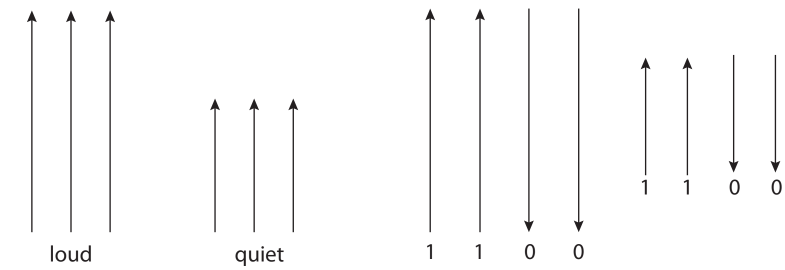


Figure 9.14 The loudness is related to the magnetic field strength in an analogue tape but this has no relevance in a digital tape.

Over a period of time the storage medium can deteriorate or become damaged leading to a corruption of the data. We have seen that both the audio tape and the computer hard drive use magnetic fields to store data. A piece of music sounds as it does because of variations in the frequency and loudness of the sound. These are recorded on an audio tape as variations in the strength and orientation of microscopic magnets, so that in a loud sound there would be a strong field and in a quiet sound a weak field (see Figure 9.14). If the magnets deteriorate with time then the strength of the field will be smaller and hence the loudness recorded will be less. We have seen that when an ADC converts a sound to a digital signal the information about loudness and frequency is recorded as a series of 1s and 0s which are recorded by the orientation of magnets. If these magnets deteriorate, the signal will not change.

Storing text

We have seen that it is possible to convert text to binary data using ASCII code and this can be stored on a hard drive, CD or other storage device. The small size of these devices makes it possible to store large amounts of data in a small space.

Retrieving damaged data

When data is converted to digital form and stored on a CD or other device then extra information is also stored, so that it is possible for the laser reader to recognize whether some of the data has been changed or damaged.

Social implications of digital data storage

In 1830, when Carl Friedrich Gauss first produced a binary electrical signal, he would have had no idea of the far-reaching social implications of his work. Today we are in the middle of the digital revolution, with new applications of digital technology arriving in the shops each week. We still don't really have any idea of the social implications for the future; however, we can see some of the effects of these advancements over the past 20 years.

A word of warning

In 1975 the data from the *Viking* Lander Mission to Mars was stored on digital tape. Unfortunately, in recent years these were found to have deteriorated and the formatting used was unreadable. Luckily there were paper copies.

In 1986 the BBC stored data for their Doomsday Project on a laser disc. This method of storage is now obsolete.

Home entertainment

Switching from analog to digital storage for music and video has made it possible to produce cheap but high quality music and video playback machines for home entertainment. Via the internet, it is now possible to download and store music and films without payment. This has resulted in lower attendance at cinemas and a whole new branch of the law to deal with illegal copying.

Digital video has also made it possible to edit and produce high quality home videos: this is good news for families but also has negative aspects.

Personal information

Because digital storage of information takes up so little space it is now possible to keep and use every piece of information about any individual. The essays you write at school can now be stored and published when you become a famous author. Every time you send an email or do an internet search, the information can be stored and used … either for you or against you. Once information is in a digital form it can also be processed in many ways; for example it is possible to use face recognition software to recognize faces in a crowd.

Advances in science and technology often have far-reaching consequences. Who should decide whether a piece of research that might have social implications should be carried out or not?

9.2 Data capture; digital imaging

Assessment statements

14.2.1 Define *capacitance*.
14.2.2 Describe the structure of a charge-coupled device (CCD).
14.2.3 Explain how incident light causes charge to build up within a pixel.
14.2.4 Outline how the image on a CCD is digitized.
14.2.5 Define *quantum efficiency* of a pixel.
14.2.6 Define *magnification*.
14.2.7 State that two points on an object may be just resolved on a CCD if the images of the points are at least two pixels apart.
14.2.8 Discuss the effects of quantum efficiency, magnification and resolution on the quality of the processed image.
14.2.9 Describe a range of practical uses of a CCD, and list some advantages compared with the use of film.
14.2.10 Outline how the image stored in a CCD is retrieved.
14.2.11 Solve problems involving the use of CCDs.

In the previous section we were mainly concerned with the storage of music and text. In this section we will look at how images can be turned into digital form so they can be manipulated and stored. It is quite possible that many readers have never used a film camera, so before we consider digital devices we should look at the workings of a simple film camera.

The film camera

A camera uses a convex lens to focus light from an image onto a film.

Lens

A convex lens refracts the light so that rays coming from an object will cross over on the film. This is called *focusing* as shown in Figure 9.15.

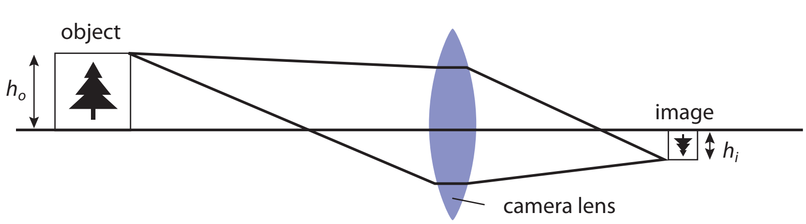


Figure 9.15 The optics of a simple film camera. If you were to look at the film you would see a small upside down version of the object, this is called an image.

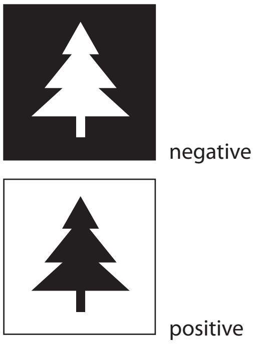


Figure 9.16

Film

A photographic film comes in a roll, which has to be loaded into the back of a camera. The film has to be inserted into the camera so that it is not exposed to light until you take a picture. The film is a plastic sheet covered with grains of a photosensitive chemical, such as silver bromide. When you take a picture, a shutter is opened and this allows light to fall on the film. When photons land on a grain, it changes the chemical nature of some of the atoms in the grain and records the image. As the film is dipped into a liquid, known as developer, the grain turns black. Since all the light areas on the image are now recorded as black dots, the picture will be a reverse of the original image. This is called a *negative* (see Figure 9.16). The developing process has to be repeated a second time to convert the negative to an image that looks like the object. This whole procedure takes some time to be performed and needs specialist equipment. Anyone who has used a digital camera will already see some of the advantages.

The digital camera

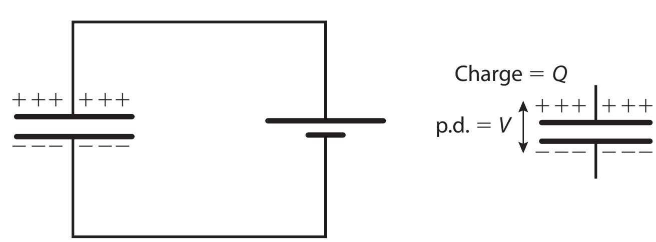
A digital camera showing the flash card used for digital storage. CCD was invented when it was found out that certain computer memory chips were light sensitive.

The optics of a digital camera are exactly the same as the film camera but instead of a film, the digital camera has a charge coupled device, or CCD, to record the image. This converts the image into a digital signal that can then be stored on a digital storage device such as a CD. To understand its operation, we must first run through some basic physical principles that have not been covered in the core chapters of this course.

Capacitance

A CCD is a small slice of silicon that has been divided into many tiny squares. Each square acts as a capacitor that stores charge released from the silicon when light shines on it. A capacitor is an electrical component that stores charge. In its simplest form it is made out of two parallel plates separated by a small air gap. If a p.d. is applied to the capacitor, then charge flows onto the plates, as in Figure 9.17. The charge can't flow round the circuit because of the air gap so it collects on the plates.

Figure 9.17 A capacitor charged by a battery stores charge. Even if the battery is now disconnected, the charge remains on the capacitor plates. We say the capacitor is charged.



The photodiode

In a CCD, the capacitors are not charged by a battery but store charge when light shines on them. Instead of air inside the capacitors there is silicon that has been impregnated with some impurity. When a photon of light is absorbed by a silicon atom, an electron becomes excited into a higher energy level. In silicon, this results in the electrons being able to move around the material freely and allowing them to collect on the capacitor plates. This arrangement is called a *photodiode*. Electrons are also liberated due to thermal energy – this is called *thermionic emission* and can cause problems, as we shall see later on.

Capacitance

When a capacitor is charged there is a p.d. between the plates. The size of the p.d. is proportional to the amount of stored charge.

$$\text{Capacitance} = \frac{\text{charge}}{\text{potential}}$$

$$C = \frac{Q}{V}$$

Unit: farad

Exercises

12 What will be the potential difference between the plates of a 15 μF capacitor if there is a charge of 5 μC on the plates?

13 The capacitance of the photodiodes in a typical CCD is about 100 μF. What charge will cause a p.d. of 1 V across it?

14 What is the charge of 50 electrons? If there are 50 electrons stored on a 100 nF photodiode what will the p.d. across it be?

The CCD array of photodiodes

The CCD is a two-dimensional (2D) array of photodiodes that have been created on a thin wafer of silicon. The wafer has many layers; a simplified version is shown in Figure 9.18.

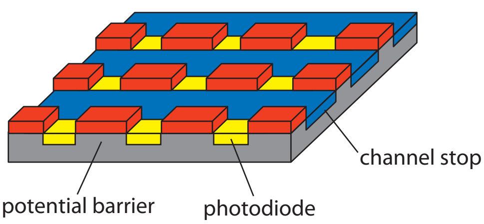
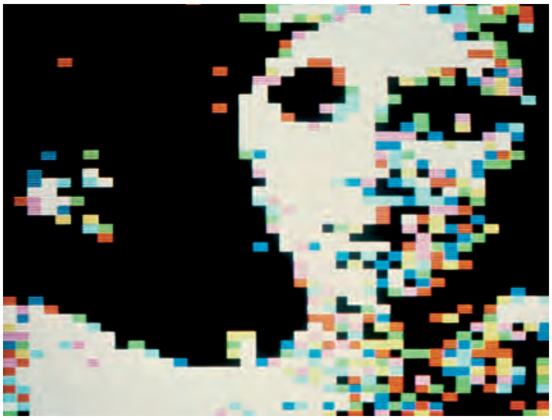


Figure 9.18 The layers of a CCD. If a photon of light lands on a photodiode, it causes an electron to be freed from a silicon atom (the photoelectric effect). This electron is trapped by the channel stop (an insulating strip) and a potential barrier created by adding impurities to the silicon. The photodiode therefore acts like a capacitor, storing charge.

Pixels

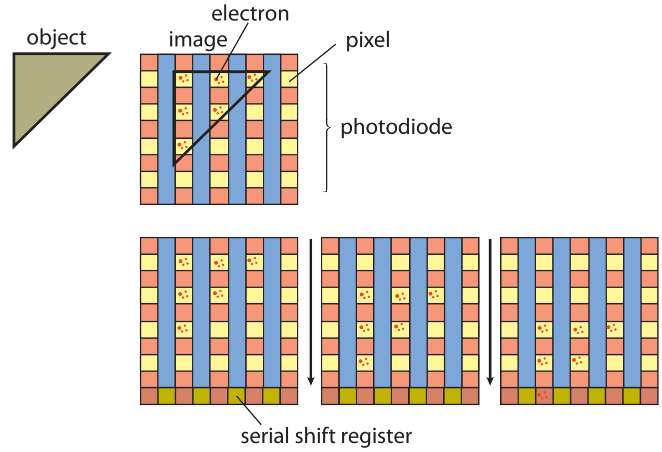
When light is focused onto the CCD by the camera lens, an image of the object is projected onto the CCD. This causes electrons to be stored in each photodiode; the number of electrons stored depends on the intensity of the light. In this way the image is recorded. However, not all the information from the light can be recorded, only the light that falls on the photodiodes. The picture has been turned into millions of small squares, or *pixels*. The number of pixels on the CCD is often quoted in camera adverts; a 6-megapixel camera has 6 million photodiodes on its CCD chip.



Artists also use pixels. Here a painting of a face is made out of small squares.

Reading the data

Figure 9.19 A pixelated image of a triangle on a CCD chip. Notice how the number of electrons stored depends on the amount of light landing on the pixel. The electrons are moved down to the serial register where they are read line by line.



The data is now stored as charge on the 2D grid of the CCD. To turn this into a digital signal we need a single line of 1s and 0s. This is achieved by applying a potential difference across the chip: this moves all the electrons down one row. The end row is called the *serial register*; this is a row of photodiodes that have wires connected to them enabling the p.d. between their top and bottom to be measured. This p.d. is proportional to the amount of charge stored (remember that $Q = CV$). As each row is moved down (clocked) the p.d.s are measured and converted to a binary number by an analogue-to-digital converter. After measurement, the charge is removed from the row. The whole process is rather like counting the eggs in an egg box by tilting the box so that the eggs roll into the end row. After the eggs are counted they are removed and another row of eggs rolls down (see Figure 9.19).

Terms and quantities related to digital imaging

Magnification (M)

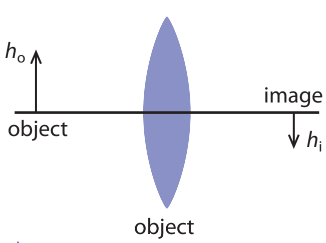


Figure 9.20

Magnification is the ratio of the height of the image and the height of the object $\left(\frac{h_i}{h_o}\right)$. If the image on the CCD is very small then only a small number of photodiodes will be illuminated, and this will result in a poor quality image. The size of the image on the CCD depends on the size of the CCD. The most common sizes of CCD range from 6 mm to 16 mm (diagonal measurement). If a large format CCD is used, then the image can be big and will be made up of more pixels.

Quantum efficiency

The quantum efficiency of a photosensitive device is defined as the ratio of the number of photons absorbed to the number of electrons freed, as a percentage. If 5 photons hit a photodiode and 3 electrons are freed then:

$$\text{quantum efficiency} = \tfrac{3}{5} \times 100\% = 60\%$$

If quantum efficiency is low, then some parts of the image will be lost.

Lenslets

If light falls on the gaps between the photodiodes, the photons will not cause any emission of electrons. To make all the photons land in the right area, small lenses are attached on top of each photodiode.

Exercises

15 What is the height of a 6 mm CCD?
If each pixel is a square of 10 μm side, how many pixels will a 6 mm CCD have?

16 A picture of a 10 m high tree is to be taken with a camera containing a 6 mm CCD. What is the magnification of the camera if the tree just fits on the CCD?

17 **(a)** If 10^{10} photons enter the camera and land on the CCD of a 5 megapixel camera, how many photons land on each photodiode?
(b) If the quantum efficiency is 70%, how many electrons are liberated in each photodiode?

18 If the rate of reading each pixel in the serial register is 5 MHz, how long will it take to read all the pixels in a 1 megapixel CCD?

19 A video camera takes 30 pictures per second. If rate of reading pixels is 5 MHz, how many pixels can the camera have?

Resolution

Resolution of an image defines the amount of detail that an image contains. You can see someone surfing in the first photo, but if the image is enlarged the person is bigger but you can't see much detail. This is because the resolution is not good enough.

A surfing scene taken on a 5 megapixel camera and an enlarged view of the surfer showing the limit of resolution. The words on the arm of the wetsuit say 'second skin' but the resolution is not good enough for you to read them.

The final image quality is determined by a combination of magnification, size of CCD and quantum efficiency.

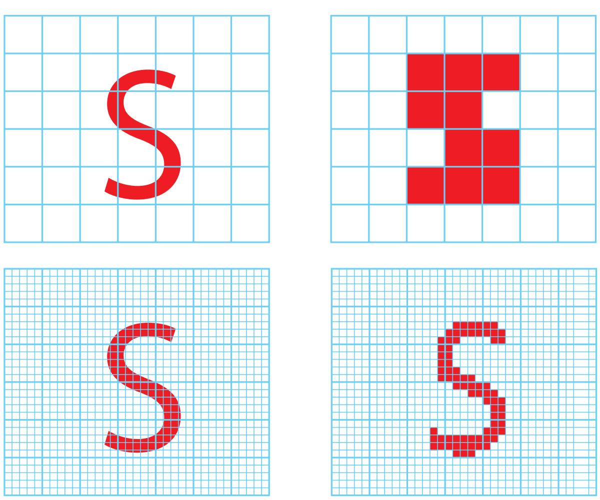


Figure 9.21 The letter S represented by 12 pixels and the same letter represented by 170 pixels.

Magnification

If the final image is very small, then all the points will be very close together. This means that it is possible for two points on the image to fall on the same photogate. These will not be distinguishable on the final picture. A large CCD will mean that the magnification can be greater, resulting in a higher resolution image.

Quantum efficiency

If the quantum efficiency is low, then not all of the incident photons will be recorded. This will result in a loss of detail, particularly in the dark areas. The quantum efficiency is different for different colours, leading to some coloured parts of objects appearing less bright.

Colour

A colour filter is fixed over the CCD in order to record colour. The pixels are grouped in three for the primary colours (red, blue and green). These three can be thought of as one pixel and they record the intensity of these three different colours of light. From this information it is possible to calculate the colour of this area of the image.

Video

A video digital camera also uses a CCD to turn the image into a digital signal. To get a moving image, the camera must take around 30 pictures per second. This means that the information from the CCD must be read very quickly. This limits the number of pixels possible and will affect the resolution, as pixels are grouped into three for a colour image. To achieve a high resolution colour image, three CCDs are used, one for each colour. This is achieved by splitting the light beam into three directions as it passes through the camera and passing each through a different colour filter. Each CCD records a different colour and they are then put together to make the final image.

Comparing digital photography with film

When digital photography was first introduced the quality of pictures was very poor. This was due to the large size of the photodiodes compared to the grain size of the photographic film. However, improvements in the CCD has meant that the photodiode size is comparable to the grain size, resulting in images with high resolution, similar to those achieved with a traditional camera. In fact, due to the much better quantum efficiency of the CCD (70%) compared to film grains (2%) the resolution can be much better.

The most obvious advantage of digital photography is the instant results; no more waiting for photographs to return from the printers. Another advantage of using a digital camera is that you can delete pictures that you don't want to print or keep. This could also be seen as a disadvantage, as the tendency is to delete all but the best records, whereas the chances of a wider variety of photographs is more likely from records of the past.

As with all digital signals, a big advantage of digital photography is the possibility of being able to use a computer to manipulate the data. Using programs, such as Adobe Photoshop, it is possible to change faces, remove buildings, and even see what people will look like when they are old.

Other uses of CCDs

The main use for CCDs is in digital photography and video, however there are many other possible applications.

X-ray machines

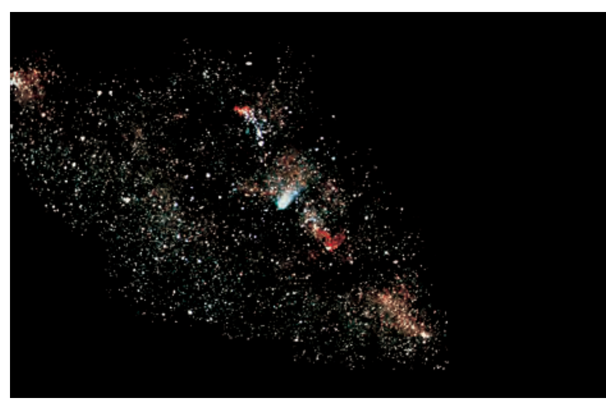
Since the photodiodes are sensitive to X-rays as well as light, they can be used in X-ray machines, thus replacing the need for film that has to be processed. An advantage of the digital image is that it can be sent to different parts of the hospital and even enhanced with colour. If you have had an X-ray recently it may well have been digital; however, the machine may have used a reusable phosphor plate that has to be scanned by a reader rather than a CCD.

Astronomy

CCDs are now commonly attached to the eyepiece of telescopes to replace the traditional film camera. This is particularly useful with remote telescopes, such as the Hubble telescope that has to send its pictures to astronomers on the Earth. The digital signal is rather easier to send than a roll of film. The CCD of a camera used in astronomy must have high quantum efficiency.

One problem with using the CCD in applications where light intensity is very low is that at normal room temperatures electrons can be excited by thermal energy. This means that even if you take a picture with the lens cap on, some electrons will be excited and the picture won't be completely black. This is not a problem when using the camera in normal daylight but can cause problems for astronomers. However, there are two ways that this can be solved. One is to cool the CCDs down to around −80 °C, and another is to remove this background light digitally. This can be simply achieved by subtracting the data collected from a photograph taken without light from the photograph of the stars.

A digital photograph from a telescope alongside the same stars but with lower quantum efficiency. The less bright stars have disappeared.



Practice questions

1 The wavelength of light used in a CD reader is 600 nm (in plastic). The 'depth' of a pit must be:

A 150 nm

B 300 nm

C 450 nm

D 600 nm

2 The binary equivalent of the number 11 is

A 1000

B 1100

C 1011

D 1110

3 If 'IB physics' were written in ASCII code the number of bits of information would be

A 9

B 10

C 72

D 80

4 The maximum number of voltage levels that can be converted into 3-bit binary are:

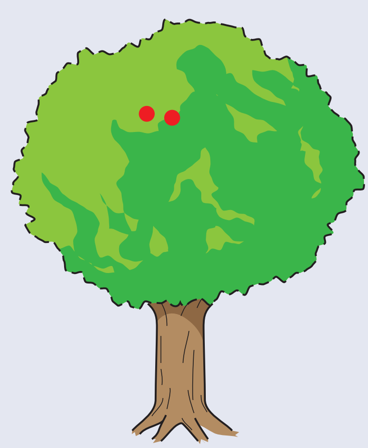
A 3

B 5

C 8

D 16

5 **(a)** A digital camera contains a 3 mm $\times$ 3 mm CCD that has 9 megapixels. It is used to take a picture of a 6 m high tree shown in the picture. The camera is focused so that image of the tree just fits onto the CCD. On the tree are two red berries separated by a distance of 2 cm. Show that the berries will be resolved in the picture.



(b) When the CCD is exposed to the light from the tree, one of the pixels receives 5×10^3 photons. Each pixel has a quantum efficiency of 80% and a capacitance of 45 pF.

(i) How many electrons will be liberated in this pixel?

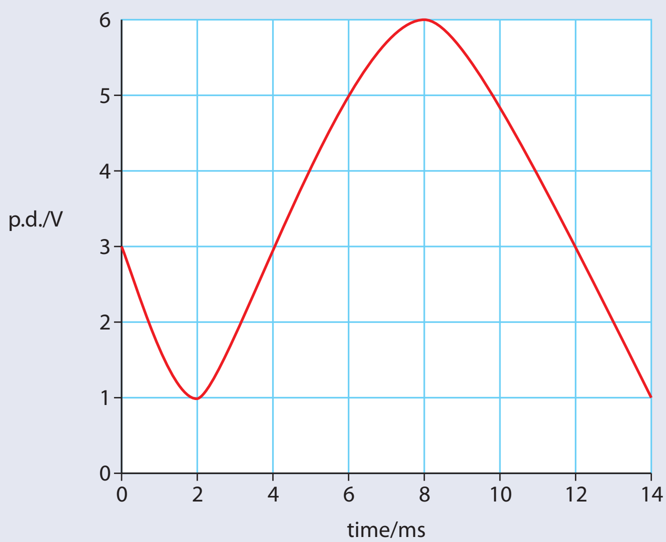
(ii) What is the p.d. across this pixel?

6 A digital recorder samples 16 bits at a rate of 44 000 Hz.

(i) Explain what is meant by '16 bits at 44 000 Hz'.

(ii) If the recorder is used to record 5 minutes of music in stereo, how many bits of information will be stored?

7 **(a)** The graph below shows an analogue signal.



This signal is sampled at 500 Hz. Fill in the table below giving the voltage level and equivalent 3-bit binary code each 2 ms.

Time/ms	Voltage level	Binary
0		
2		
4		
6		
8		
10		
12		

(b) Give one advantage of storing music in digital rather than analogue form.

10 Astrophysics (Option E)

10.1 Introduction to the universe

Assessment statements

E.1.1 Outline the general structure of the solar system.
E.1.2 Distinguish between a stellar cluster and a constellation.
E.1.3 Define the light year.
E.1.4 Compare the relative distances between stars within a galaxy and between galaxies, in terms of order of magnitude.
E.1.5 Describe the apparent motion of the stars/constellations over a period of a night and over a period of a year, and explain these observations in terms of the rotation and revolution of the Earth.

The view from here

We know that the stars have been of interest to mankind since the time when we lived in caves; this interest probably arose from the fact that we could make predictions based on things that happen in the sky. When the Sun comes up, it gets light, and when it goes down, it gets dark. If people counted the sun-ups and sundowns, they could predict when flowers would bloom and fruit ripen, and from the position of the Moon they could calculate the tides. Since the largest objects in the sky gave such good predictions, it was a small step to assume that all the stars could be used to predict events too. Ancient civilizations were so convinced about this that they built huge structures to aid their calculations and called these bright bodies gods.

Stonehenge, built to predict when eclipses would take place.

As people began to travel they would have noticed that the motion of these bodies depended on where they were. At the equator, the Sun's path does not change very much from day to day, but close to the North and South Poles, the Sun doesn't go down in the summer but describes a big circle in the sky. In the winter, it isn't visible at all. The stars also have different paths; on the equator, stars come up at 90° to the horizon but at the poles they travel parallel to the horizon.

You might have noticed that although the stars move across the sky, their relative position doesn't change. If you join the dots, you can make the same patterns every day; you may know the name of some like the Big Dipper or Orion. If you watch the sky for several days you will notice that some of the 'stars' don't stay in the same place (there are about six that you can see without a telescope). They wander east and west. This strange behaviour didn't go unnoticed by the ancient civilizations, which decided that these must be gods – we now call them planets.

The path of stars in the northern hemisphere.

When looking at the stars, it appears that either they are rotating around the Earth or the Earth is spinning. The ancients believed in the first explanation but we now know it is the second. The motion of the planets is much more difficult to work out.

Explaining the movement

The movements of the stars, Sun, planets and Moon through the sky are complicated by the fact that the Earth, Moon and planets are all moving relative to each other. So to simplify things we will take one at a time.

Movement of the stars

The movement of the stars is due to the rotation of the Earth, but what you see depends where on the Earth you stand. If you stand at the North Pole in the middle of the winter you get a very clear picture because the Sun doesn't rise – so you can see the stars all the time. You would see the stars moving round in the same way as you would see objects in your room move if you sat on a spinning chair. Stars on the horizon move horizontally, but if you look up they make circles with the pole star in the middle. If you time this rotation, you would find that one complete revolution always takes the same time, 23 hours 56 minutes. This is the time period for the Earth's rotation and it gives us our first complication, as it's not the same as our day. This is because the length of a day is based on the Sun not the stars. The effect is that if you note where the stars are at 12:00 each day they will have moved a little bit further forward. This is illustrated in Figure 10.1.

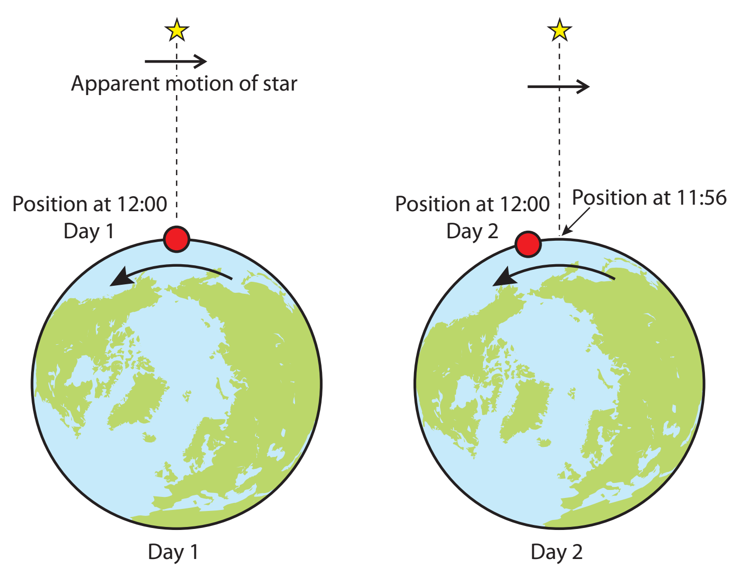


Figure 10.1 The position of an observer on the Earth is slightly different from one day to the next, causing the stars to move slightly to the west each day. East and west are confusing but if you find your country on the map then think how it would appear from there.

To access some simulations showing positions of stars and the Sun, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblinks 10.1 and 10.2.

If time for the Earth to rotate through 360° is 23 hours and 56 minutes, in 4 minutes it will rotate through 1°. This means that in 360 days you will be looking at the stars in the same position.

The movement of the Sun

As the Earth spins on its own axis, it is also rotating around the Sun. This doesn't affect the position of the stars because this movement is very small compared to the distance to the stars, but it does affect when we can see them. If the axes of the Earth and the Sun were the same, then the Sun would simply appear and disappear at the same place every day, as the stars do. The time at which it appeared would be slightly different due to the fact that we are moving around the Sun at the same time as we rotate on our axis. But the two axes of rotation are different, causing the Sun to move in a north-south direction from day to day, as illustrated in Figure 10.2. Note in this diagram the pole star would be a long way above the top of the diagram; this and the other stars are our fixed reference positions.

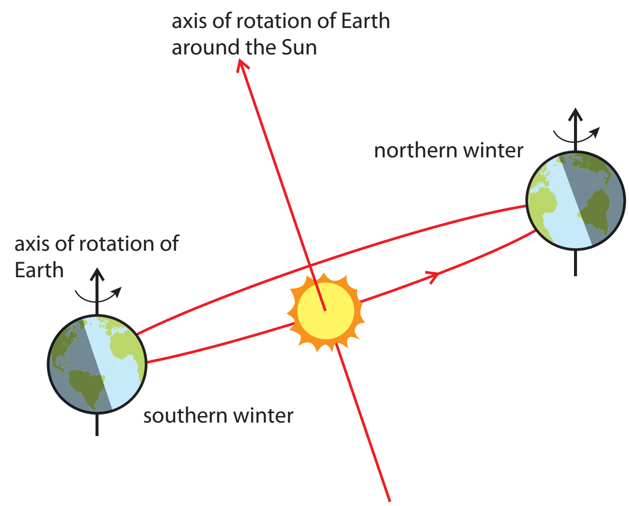


Figure 10.2 The axis of rotation of the Earth's rotation and its orbit around the Sun are not the same. This is why during the winter in Europe it appears more towards the south than in the summer.

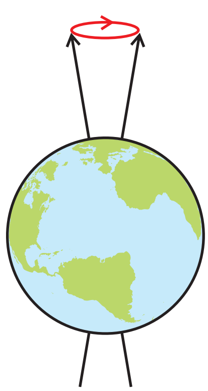


Figure 10.3 The precession of the Earth's axis.

Precession

The Earth is not a perfect sphere; it's squashed so that it is fatter round the equator than at the poles. This means that when it is in one of the positions shown in Figure 10.2, the force of gravity exerted by the Sun on the Earth will cause a turning effect, pulling the axis of rotation slightly out of line. This effect is called *precession* and it causes the axis to rotate as shown in Figure 10.3. This makes all the stars, including the pole star, appear to move. However, since the period of rotation is 26 000 years, it is not very noticeable.

Movement of the planets

The planets also orbit the Sun. This is very simple to understand when visualizing the solar system from a great distance, but not so easy to picture when standing on the Earth. From the Earth we see the planets wandering back and forth from one night to the next, sometimes moving east and sometimes west; this is why they were given the name *planet*, from the Greek for *wanderer*. Figure 10.4 illustrates how this motion can be the result of planets orbiting the Sun.

To see an animated version of the planetary motion, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 10.3.

To further complicate matters, the planet orbits are not in exactly the same plane as the Earth's, resulting in a north-south movement when viewed from Earth.

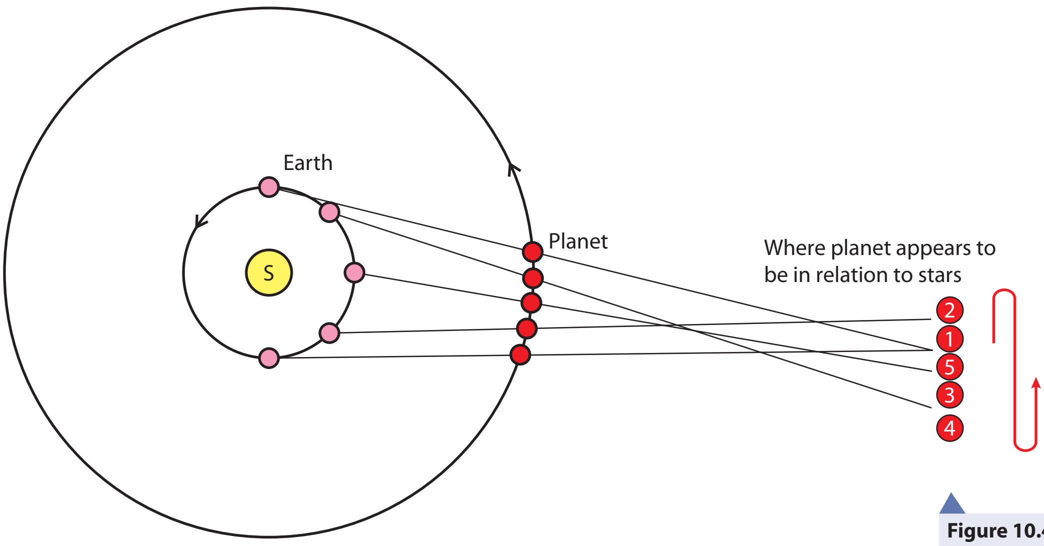
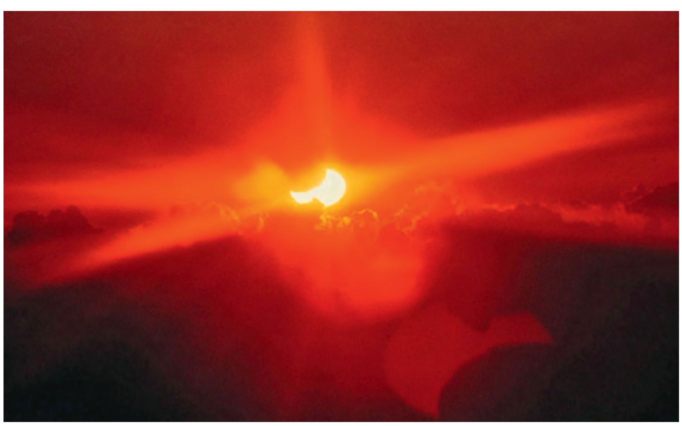


Figure 10.4 A planet orbiting more slowly than the Earth would sometimes appear to advance and sometimes fall behind when viewed from Earth.

The Moon

The Moon orbits the Earth once every 27.3 days, causing it to appear at different times each day. The Moon is not only seen at night, but sometimes appears ghostlike in the daytime sky. When the Moon is in front of the Sun, an eclipse takes place.



A partial eclipse of the Sun.

The solar system

The solar system is the name given to everything that orbits the Sun, including the planets and their moons, asteroids and comets.

When modelling gravity, we treated orbits as circular for simplicity, but in fact, the planets have slightly elliptical orbits. An ellipse is a flattened circle with two centres; one of these centres is the Sun. We know that for a satellite to have a circular orbit at a given radius, it must have a very specific velocity. If it goes faster, its orbit will be elliptical or hyperbolic.

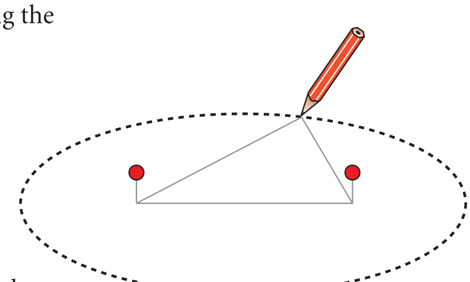


Figure 10.5 Drawing an ellipse with pins and a piece of string.

The solar system was formed from a spinning cloud of dust. The dust clumped together to form planets and asteroids that we see today. It would have been very unlikely for those clumps to have exactly the right speed for circular orbits, so they are ellipses. In an elliptical orbit, the distance between the planet and the Sun is not constant; this means that its PE is not constant. As it moves away, its PE will increase. Since energy must be conserved, this means that the KE of the planet must decrease, resulting in a change of speed.

The planets

On August 24th, 2006 the International Astronomical Union (IAU) declared the official definition of a planet:

> A 'planet' is a celestial body that (a) is in orbit around the Sun, (b) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a hydrostatic equilibrium (nearly round) shape, and (c) has cleared the neighbourhood around its orbit.

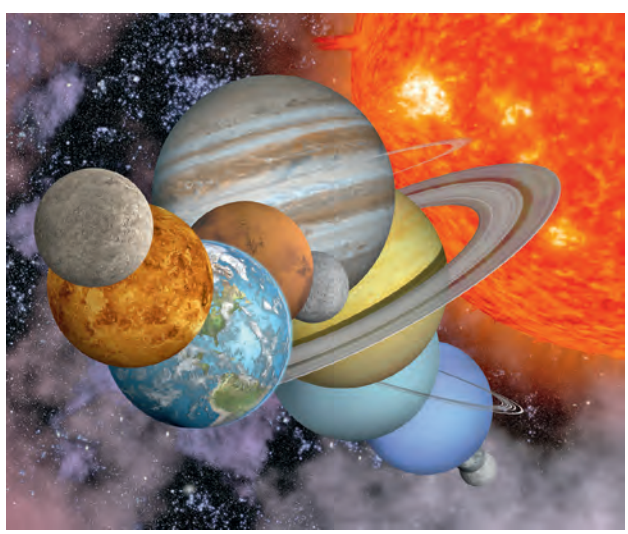
There are 8 planets in orbit around the Sun. Each planet has a different radius, time period and size.

Their order in terms of distance from the Sun is:

Planet	Orbit radius (m)	Mass (kg)	Radius (m)	Period
Mercury	5.79×10^{10}	3.3×10^{23}	2.44×10^{6}	88.0 days
Venus	1.08×10^{11}	4.87×10^{24}	6.05×10^{6}	224.7 days
Earth	1.50×10^{11}	5.98×10^{24}	6.38×10^{6}	365.3 days
Mars	2.28×10^{11}	6.42×10^{23}	3.40×10^{6}	687.0 days
Jupiter	7.78×10^{11}	1.90×10^{27}	6.91×10^{7}	11.86 years
Saturn	1.43×10^{12}	5.69×10^{26}	6.03×10^{7}	29.42 years
Uranus	2.88×10^{12}	8.66×10^{25}	2.56×10^{7}	83.75 years
Neptune	4.50×10^{12}	1.03×10^{26}	2.48×10^{7}	163.7 years

The relative sizes of each are illustrated in the photo.

The planets and the Sun drawn to scale (the size is to scale, not the orbit).



There are also over 40 dwarf planets of which Pluto is one. The reason Pluto is not a planet is because it does not dominate its neighbourhood; in fact Pluto is only twice as big as its moon Charon.

Exercise

1 Using your knowledge of gravity, calculate the period of the planets from their radius (assume the orbits are circular). You could do this in a spreadsheet for convenience.

Asteroids

Between the orbits of Mars and Jupiter there are billions of smaller orbiting bodies called asteroids (as seen in the photo above). These range from the size of dust particles to several hundred kilometres in radius.

Comets

A comet is also a small orbiting body, but unlike an asteroid it is made up of loose particles of ice and rock that are blown off by the solar wind, forming a tail. One of the most famous comets, Halley's comet, has a period of 76 years and a very elongated orbit. Other comets have hyperbolic orbits, passing only once.

Halley's comet photographed in 1986.

Beyond the solar system

The reason we can see the planets is not because they give out light, but because they reflect the light from the Sun. There are certainly countless other cold bodies (big and small) as we go beyond the solar system, but we can't see them since they are too far away. All the other objects we see in the sky must therefore give out light, and these are the stars.

Stars

Stars come in all shapes, sizes and colours and have names to match, such as *white dwarf*, *red giant* and *supernova*. They are massive balls of plasma (a gas in which the particles are moving so fast that the electrons have all been knocked off the atoms and move about freely). Since they are the bodies that we can see best they are the ones that we know most about; most of this chapter will be about stars.

Different cultures developed different systems for defining the different patterns in the stars. The Greeks had the 12 signs of the zodiac, the Chinese 28 xiu and the Indians 27 Nakshatra.

Constellations

Constellations are the results of ancient civilisations playing 'join the dots' with the stars. The stars are not related by anything physical except maybe that they are all bright.

The Orion constellation – a man with big shoulders and a small head? The three stars in the middle are supposed to be a belt.

Stellar cluster

A stellar cluster is a group of stars that are physically close together rather than looking as though they are. These are formed by the collapse of a gas cloud.

Units of length

Astronomical unit AU

1 AU = 1.5×10^{11} m

This is the distance between the Sun and the Earth.

Light year (ly)

1 light year = 9.46×10^{15} m

This is equal to the distance travelled by light in one year.

Parsec (pc)

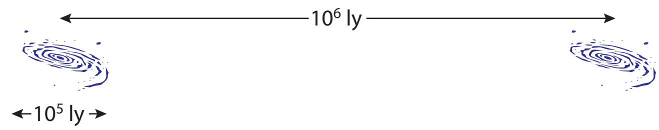
1 parsec = 3.26 ly

This is defined by making a triangle between the Earth, the Sun and a distant object . If the angle at the distant object is l arcsec then it would be 1 parsec away. More about this later.

Galaxies

A galaxy is a large collection of stars held together by gravity. A galaxy can contain trillions of stars and are between 10^3 and 10^5 light years across. Each star is approximately 1 ly apart and the galaxies are about 10^6 ly apart. The relative sizes are illustrated in Figure 10.6.

Figure 10.6 The distance between galaxies is about 10× their size. The stars are typically 1 ly apart.



We belong to a galaxy called the Milky Way. If you look into the sky at night you can see a vague stripe where the stars are a bit denser; this is the Milky Way. The shape of our galaxy is a flat spiral, like the ones in Figure 10.6, and we live near to the edge. The stripe in the sky is the cross-section of the spiral. Other galaxies can be round, lens shaped or irregular.

The Milky Way galaxy.

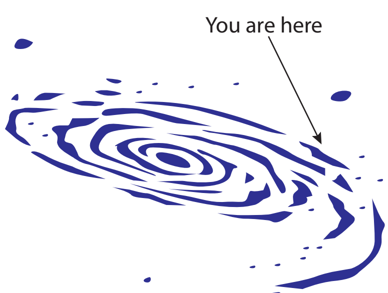


Figure 10.6 The Milky Way showing the position of the Sun.

Exercises

2 The distance to the nearest star is 4×10^{13} km. What is this in light years?

3 How long does it take light to travel from the Sun to the Earth?

4 How long would it take for a rocket travelling at 30 000 km h^{-1} to travel to the nearest star from the Earth?

10.2 Stellar radiation and stellar types

Energy source of the stars

Assessment statements

E.2.1 State that fusion is the main energy source of stars.

E.2.2 Explain that, in a stable star (for example, our Sun), there is an equilibrium between radiation pressure and gravitational pressure.

The closest star to us is the Sun so this is the star we know most about. By calculating the amount of energy that the Sun releases and knowing the size and mass of the Sun, it is possible to work out the temperature of the inside. We find that the conditions inside are just right for the fusion of nuclei to be the source of energy. On the Sun it is the fusion of hydrogen into helium that provides the energy, but this isn't the same for all stars. This reaction takes place inside the Sun so it is difficult to obtain direct evidence that fusion is taking place. However, we know that this reaction produces neutrinos and that they would pass through the outer layers of the Sun and travel to the Earth. Neutrinos are difficult to detect but are present in the radiation from the Sun.

The neutrino problem
For many years the number of neutrinos detected on Earth did not match the number predicted. This was resolved by Masatoshi Koshiba and Raymond Davis Jr. who in 2002, shared half of the physics Nobel prize for their work. The discrepancy was due to neutrinos changing their type, showing that neutrinos must have mass.

The proton-proton chain

The fusion process taking place on the Sun is called the proton-proton chain. It is not possible to simply join two protons and two neutrons together to form helium, because momentum and energy cannot be conserved in this process. The chain of events to get from 4 hydrogens to 1 helium is shown in Figure 10.7

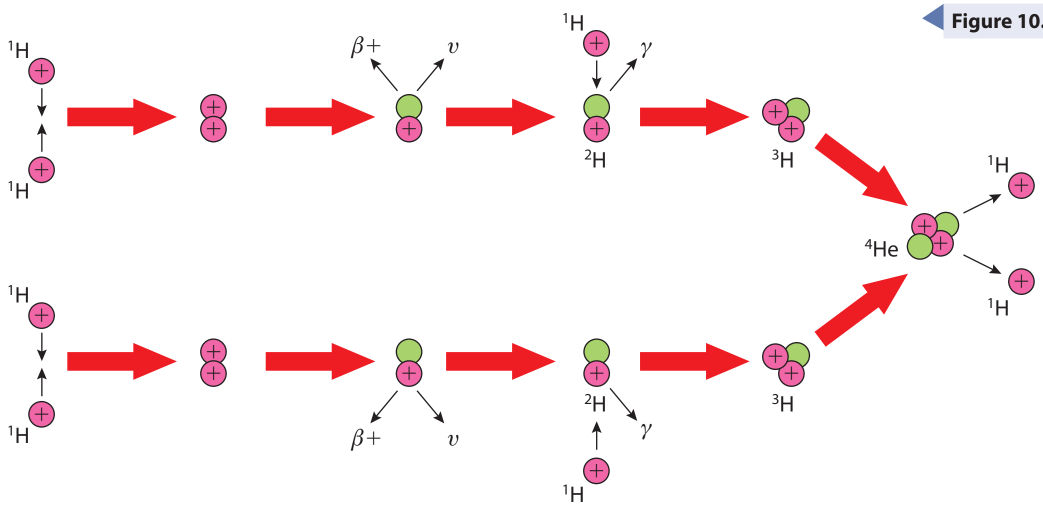


Figure 10.7 The proton-proton chain.

This is a complicated reaction but can be summarized in the following equation:

$$4\,^1_1\text{H} = \,^4_2\text{He} + 2e^+ + 2v_e + 2\gamma$$

This reaction releases energy due to the fact that the mass of the products is less than the mass of the original hydrogen nuclei; this mass is converted to energy as explained in Chapter 7. The amount of energy per reaction is 26.7 MeV.

Stable stars

The Sun is a stable star. This means that it isn't getting bigger or smaller. The mass of the Sun is so large that the gravitational force experienced by the outer layers is very big; this alone would cause the Sun to collapse. The reason it doesn't collapse is due to the continual production of energy from the fusion reaction in its core. This gives the particles KE causing a pressure (radiation pressure) that pushes back against gravity. We can compare this to a balloon; the rubber of the balloon pushes in and the gas pressure pushes out. A balloon doesn't need a source of energy because its temperature is the same as the surroundings but if it lost heat, the inside pressure would drop and the balloon would shrink. Since the Sun is losing energy all the time, it needs a source of energy to remain stable.

10.3 Light from stars

Assessment statements

- E.2.3 Define the luminosity of a star.
- E.2.4 Define apparent brightness and state how it is measured.
- E.2.5 Apply the Stefan–Boltzmann law to compare the luminosities of different stars.
- E.2.6 State Wien's (displacement) law and apply it to explain the connection between the colour and temperature of stars.
- E.2.7 Explain how atomic spectra may be used to deduce chemical and physical data for stars.
- E.2.8 Describe the overall classification system of spectral classes.

Early measurement
Before techniques had been invented for measuring brightness, it was estimated with the naked eye on a scale from 1 to 6. The brightest stars were 1 and the ones that you could only just see were 6. This would be dependent on how good your eyesight was.

The reason we can see stars is that they give out light. By measuring this light we can gain a lot of information about the temperature, size and chemical composition of the star.

Luminosity (*L*)

The luminosity of a star is defined as:

the total amount of energy emitted by the star per second.

The unit is the watt.

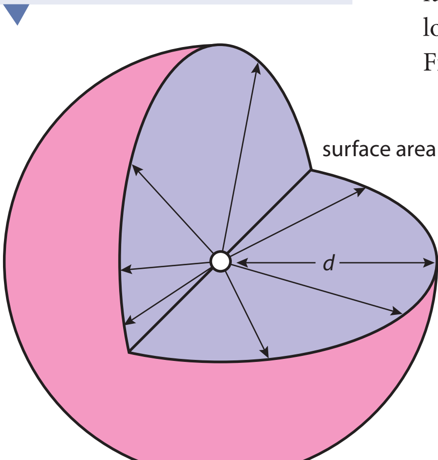
The luminosity depends upon the temperature of the star and its size. If two different size stars have the same temperature, the bigger one will give out more energy than the small one.

The Sun has a luminosity ($L_\odot$)of 3.839×10^{26} W, and a star's luminosity is sometimes quoted as a fraction or multiple of this.

Apparent brightness (*b*)

If you look into the sky at night, you will see that some stars are bright and some dim. The brightness depends on how much light is entering your eye, and this depends upon how much light the star gives out (its luminosity) and how far away it is. If you look at two identical stars, one close and the other far, the closer one looks brighter because the light hasn't spread out so much. This is illustrated in Figure 10.8. The definition of apparent brightness is:

Figure 10.8 As the light travels away from the star, the energy is spread over a bigger area.



the amount of energy per second received per unit area.

The unit is $\mathrm{W\,m^{-2}}$.

The apparent brightness is related to the luminosity by the equation

$$b = \frac{L}{4\pi d^2}$$

where d = the distance to the star.

To calculate the star's brightness, we must measure the energy per unit second absorbed by a detector placed perpendicular to the direction of the star. This can be done by using a telescope to focus an image of the star onto the CCD plate of a digital camera.

From the p.d. across each affected photodiode it is possible to calculate how much energy per square metre is incident at the Earth.

Exercises

5 The luminosity of the Sun is 3.839×10^{26} W and its distance from the Earth is 1.5×10^{11} m. Calculate its
(a) apparent brightness
(b) brightness at a distance of 10 pc.

6 Sirius, the brightest star, has a luminosity 25 times greater than the Sun and is 8.61 light years from the Earth. Calculate
(a) its apparent brightness
(b) its brightness at a distance of 10 pc.

7 If the luminosity of a star is 5.0×10^{31} W and its apparent brightness is $1.4 \times 10^{-9}\,\mathrm{W\,m^{-2}}$, calculate its distance from the Earth in ly.

The Sun $\odot$
$\odot$ is used as the symbol for the Sun
so $M_\odot$ is the mass of the Sun
$L_\odot$ is the luminosity of the Sun.

Black body radiation

A hot body emits light due to electron energy transitions. If the temperature is increased, the amount of energy available increases, so the electrons can gain more energy and move into higher energy levels. The result is that not only are more photons released, but their average energy is greater. Since $E = hf$, higher energy means higher frequency or shorter wavelength. This changes the spectrum of light as illustrated in Figure 10.9.

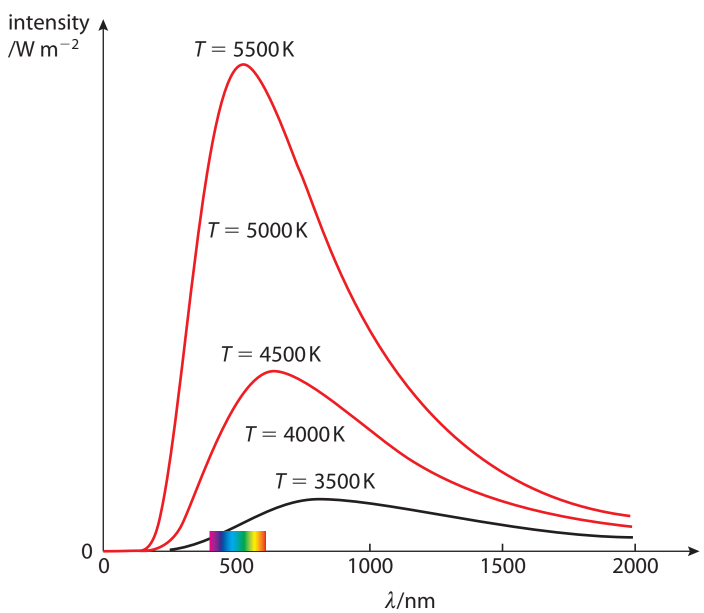


Figure 10.9 The intensity distribution for bodies of different temperature, the spectrum indicates where the visible region lies on the scale.

Black body
A black body absorbs all wavelengths of light and reflects none. It is also a perfect emitter of radiation. Stars are approximately black body radiators.

We can think of this as a bar chart, with each bar representing the intensity of radiation at different wavelengths. The total intensity is therefore the sum of all the bars, which is the area under the curve. This can be found using **the Stefan-Boltzmann law**

$$\text{power per unit area} = \sigma T^4$$

where $\sigma = 5.6 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$ (the Stefan-Boltzmann constant).

If a star has a surface area A and temperature T then the total power emitted (the luminosity), L is given by:

$$L = \sigma A T^4$$

We can see from the graph that as the temperature increases, the wavelength of the peak gets less. The relationship between the peak wavelength and temperature is given by the **Wien displacement law**

$$\lambda_{max} = \frac{2.90 \times 10^{-3}\,\text{km}}{T}$$

Worked example

The maximum in the black body spectrum of the light emitted from the Sun is at 480 nm. Calculate the temperature of the Sun and the power emitted per square metre.

Solution

Using Wien's law

$$\lambda_{max} = \frac{2.90 \times 10^{-3}}{T}$$

$$T = \frac{2.9 \times 10^{-3}}{\lambda_{max}} = \frac{2.9 \times 10^{-3}}{480 \times 10^{-9}} = 6000\,\text{K}$$

Now using the Stefan-Boltzmann law

Power per unit area $= 5.6 \times 10^{-8} \times (6000)^4 = 7.3 \times 10^7\,\text{W m}^{-2}$

If the radius of the Sun is 7.0×10^8 m, what is the luminosity?

The surface area of the Sun $= 4\pi r^2 = 6.2 \times 10^{18}\,\text{m}^2$

The total power radiated $= 6.2 \times 10^{18} \times 7.3 \times 10^7 = 4.5 \times 10^{26}\,\text{W}$

Exercises

8 The star Betelgeuse has a radius of 3.1×10^{11} m and a surface temperature of 2800 K. Find its luminosity.

9 The intensity peak in a star's spectrum occurs at 400 nm. Calculate
(a) its surface temperature
(b) the power radiated per square metre.

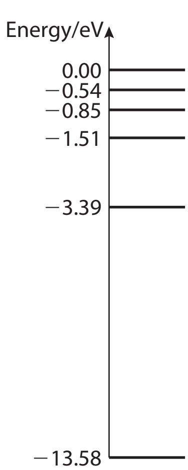


Figure 10.10 The energy levels of hydrogen.

Stellar spectra

We have seen that a star emits a continuous spectrum of electromagnetic radiation with a peak in intensity that is dependent on its temperature. As this light passes through the outer layers of the star some of it is absorbed.

Absorption spectra

You might remember from Chapter 7 that the electrons in atoms only exist in a certain energy level. That means that when they are excited they give light that has discrete wavelengths, resulting in a line spectrum. If white light is passed through the same gas, these same wavelengths are absorbed, leaving dark lines in the otherwise continuous white light spectra. This is what happens when the black body radiation from the star passes through the outer layers. Since the lines are different for each element, it is possible to determine which elements are present by analysing the lines.

Figure 10.11 The spectrum of a star showing clear absorption lines for hydrogen.

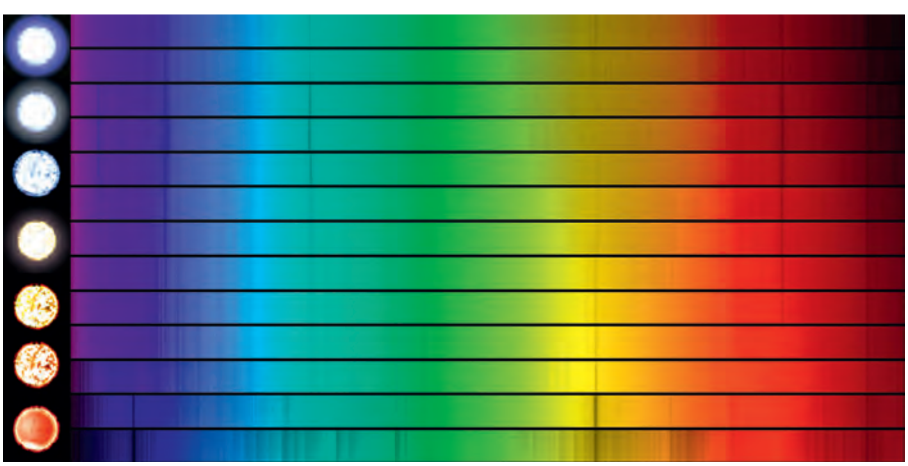
In addition it is possible to calculate the temperature of the gas; let us take hydrogen as an example (see Figure 10.10). When the gas is hot, most of the electrons will already be in the higher energy levels, so when they absorb radiation they cannot make the biggest jump (from −13.58 to 0.00) this means that the higher energy photons will not be absorbed. This results in a weak absorption line for that wavelength. Careful analysis of the relative strength of spectral lines is an accurate way of determining star temperature.

Spectral classification of stars

We have seen that the spectrum of a star is related to its temperature and chemical composition. It also determines its colour: if the peak is at the blue end, it will be blue, and if at the red end then it will be red. The Harvard classification classifies stars according to their colour, with each class assigned a letter OBAFGKM, as shown in the table below.

● **Hint:** Oh Be A Fine Girl Kiss Me (or guy or gorilla if you prefer) is a common way of remembering this unusual sequence.

Class	Temperature	Colour
O	30 000–60 000	Blue
B	10 000–30 000	Blue–white
A	7 500–10 000	White
F	6000–7500	Yellow–white
G	5000–6000	Yellow
K	3500–5000	Orange
M	2000–3500	Red



The spectra and star colour for different stars starting with O at the top ending with M at the bottom (the extra ones are sub categories). Notice the varying strength of the hydrogen absorption lines.

Doppler shift

The apparent frequency of a light source is dependent on the relative motion of source and observer. If they move towards each other, the wavelength is reduced and if they move apart it is increased. This is the Doppler effect. If a star is moving relative to the Earth, its whole spectrum will shift. This can be recognized by

It isn't just the hydrogen lines that shift – it's the whole spectrum. But the continuous part appears the same. It's only because we know that the dark lines were produced by hydrogen that we can tell that their wavelength has changed.

measuring the position of the characteristic absorption lines of, for example, hydrogen.

Red shift – longer λ – star moving away.

Blue shift – shorter λ – star moving closer.

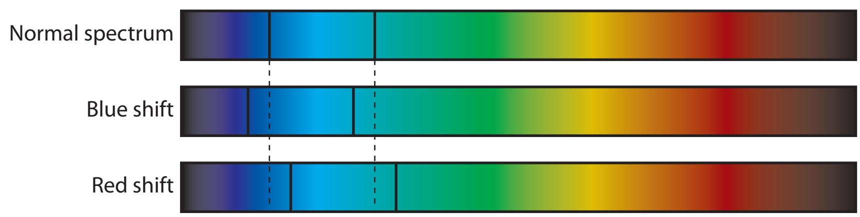


Figure 10.12 The spectra from 3 identical stars, one stationary and two moving relative to the Earth.

10.4 Types of star

Assessment statements

E.2.11 Identify the general regions of star types on a Hertzsprung–Russell (HR) diagram.
E.2.9 Describe the different types of star.
E.2.10 Discuss the characteristics of spectroscopic and eclipsing binary stars.

The Hertzsprung–Russell diagram

If you know the temperature of a star and want to calculate its luminosity, you would have to know how big it was, as a small hot star could give out the same amount of energy per second as a big cool one. If you plotted the luminosity and temperature of all the stars on a chart, you'd probably expect the chart to be covered in dots, since you would expect to find every type of star somewhere in the universe, small hot, big hot, small cool etc. When Ejnar Hertzsprung plotted this chart in 1905 he found that this was not the case, as shown in Figure 10.13.

From this, we can see, for example, that there are very few small hot stars and that most stars exist on a band down the middle. This pattern is related to the way stars evolve and leads to a classification of the stars into four groups.

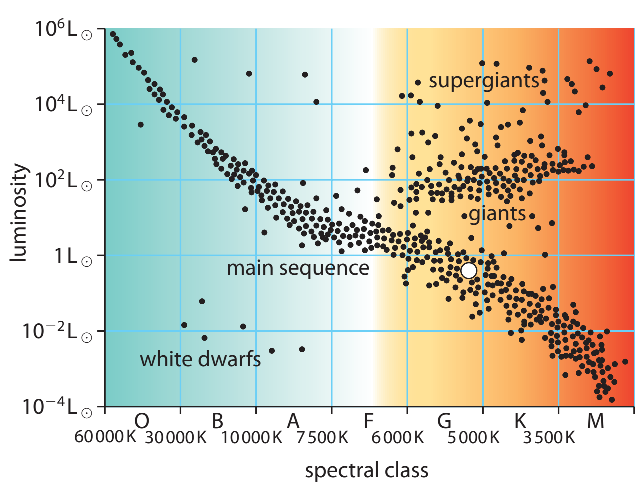


Figure 10.13 A Hertzsprung–Russell diagram. The white dot is the approximate position of the Sun. Notice that temperature increases to the left and the scales are not linear.

In physics we often draw line graphs showing the relationship between two quantities. An HR diagram isn't a line graph but a chart. Each dot represents a different star, but it would be meaningless to draw a best fit line.

Main sequence stars

90% of the stars we see at night are in the diagonal band crossing the centre of the diagram from hot and large on the left to small and cool on the right. These are called the *main sequence stars*, of which the Sun is one.

Giants

A cool star that gives out a lot of energy must be very big, so these are called giants. The coolest M class stars are called *red giants* due to their colour. The luminosity of a giant is about 100 times bigger than the Sun. If they are the same temperature as the Sun, they must have an area 100 times bigger, therefore a radius 10 times bigger. If their temperature is lower they can be even bigger.

Supergiants

A supergiant is a very big cool star. With luminosities 10^6 times greater than the Sun, they have radii up to 1000 times that of the Sun. These are very rare stars but one is very easy to spot: Betelgeuse is the right shoulder of Orion and you can see it in the photo on page 337.

White dwarfs

A white dwarf is a small hot star, hotter than the Sun but only the size of the Earth. They have a low luminosity so aren't possible to see without a telescope.

Variable stars

A variable star has a changing luminosity, so its position on the HR diagram is not constant. This is due to a change in size of the star. As it gets bigger so its luminosity increases. This variation is sometimes cyclic as in a Cepheid variable.

Calculating the mass of a star

Binary stars are pairs of stars that orbit each other. There are many of these in universe and they enable the astrophysicist to calculate the star's mass.

Remember for a satellite in a circular orbit

$$\frac{GMm}{r^2} = m\omega^2 r$$

So if we measure $\omega \left(\frac{2\pi}{\text{time period}}\right)$ and r, we can calculate M. This can be applied when the orbiting star is much smaller than the central star. Figure 10.14 shows how the orbits would be with different relative masses. If you imagine the stars to be stuck on the ends of a stick it might help you to visualize the different paths.

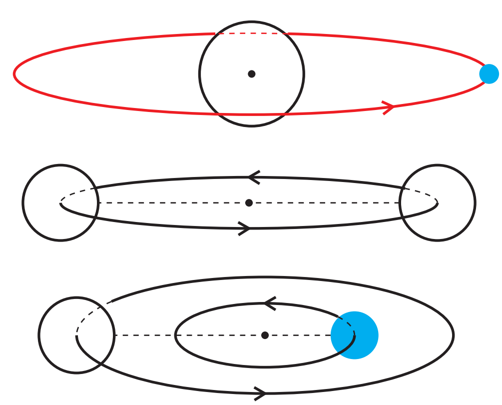


Figure 10.14

When the masses are similar, the time period is given by the formula

$$T^2 = \frac{4\pi^2 d^3}{G(M_1 + M_2)}$$

where $(M_1 + M_2)$ is the total mass and d is their separation.

By comparing the mass of these stars to their luminosity, we find $L \propto M^{3.5}$ so if we know any star's luminosity, we can find its mass.

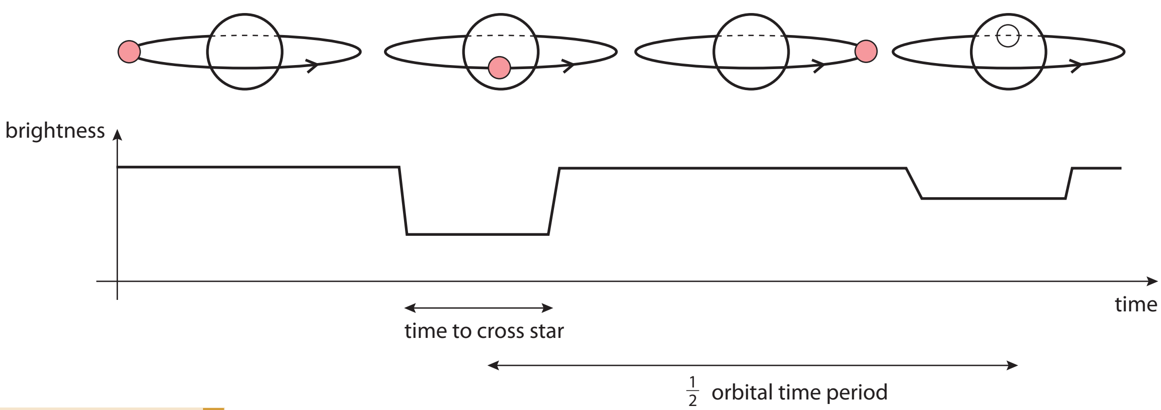
Visual binaries

These are pairs of stars that can be seen to rotate around each other; in this case it is simply a case of measuring the time period and radius directly. Other examples are too far away or too small to resolve the individual stars; we only know they are binaries due to fluctuations in the light we receive.

Eclipsing binary

If the orientation of binary stars orbit causes them to periodically pass between the Earth and each other then they eclipse each other. When this happens it causes a reduction in the stars apparent brightness, as shown in Figure 10.15.

Figure 10.15 The variation in apparent brightness for an eclipsing binary. The graph is called a light curve. Notice how the drop in brightness is bigger when the less bright star is blocking the bright one.



To see how the light curve is related to the stars, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 10.4.

From the light curve it is possible to get information about the orbital period and the relative size of the stars. Spectral analysis of the light from each star passing through the atmosphere of the other can also give information about the atmosphere of the two stars.

To see how the spectrum is related to the stars, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 10.5.

Spectroscopic binary

As the binary stars in Figure 10.16 orbit each other, they are sometimes moving towards the Earth and sometimes away. This will cause a varying Doppler shift in the light received on the Earth as shown in Figure 10.16.

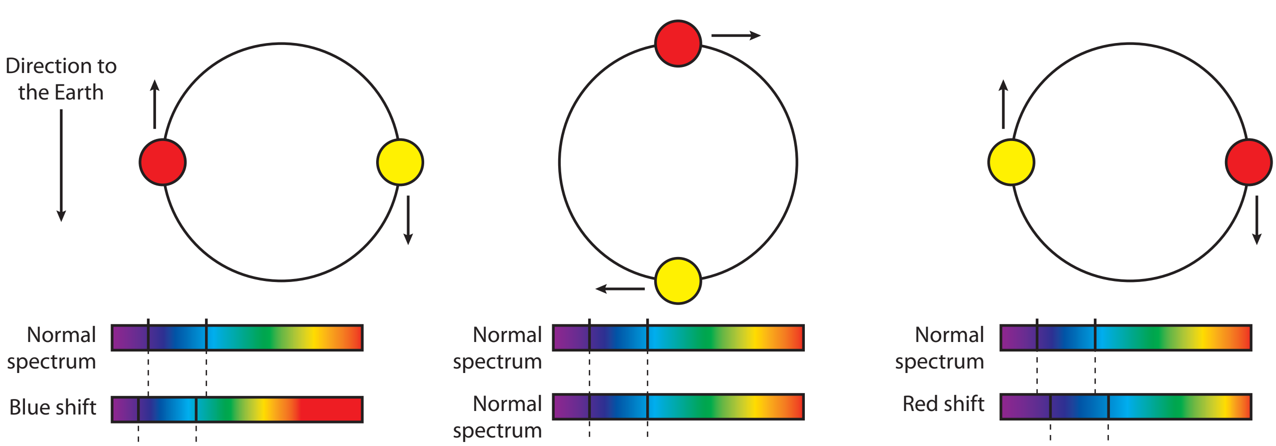


Figure 10.16 The Doppler shift for a binary star at different positions. In this example the yellow star is the only one considered since it is much brighter. If they have equal brightness then both red shift and blue shift occur resulting in two lines.

Exercises

10 With reference to the HR diagram classify the following stars:

(a) Luminosity = $L_{\odot}/100$, temperature = 30 000 K

(b) Luminosity = 100 $L_{\odot}$, temperature = 5000 K

(c) Luminosity = 10 000 $L_{\odot}$, temperature = 30 000 K

11 With reference to the light curve for binary stars shown in Figure 10.17, calculate the orbital period. What can you deduce about the orbital radius compared to the size of the stars?

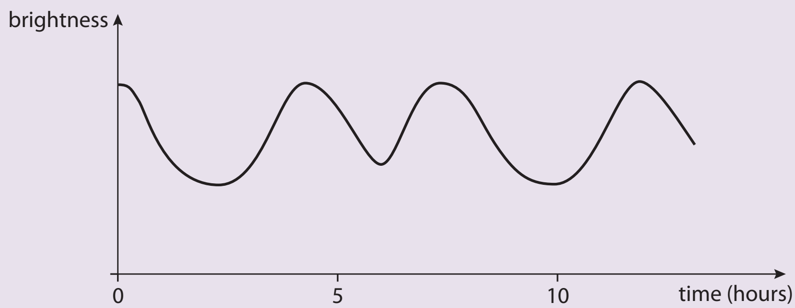


Figure 10.17

10.5 Stellar distances

Assessment statements

E.3.1 Define the *parsec*.

E.3.2 Describe the stellar parallax method of determining the distance to a star.

E.3.3 Explain why the method of stellar parallax is limited to measuring stellar distances less than several hundred parsecs.

E.3.4 Solve problems involving stellar parallax.

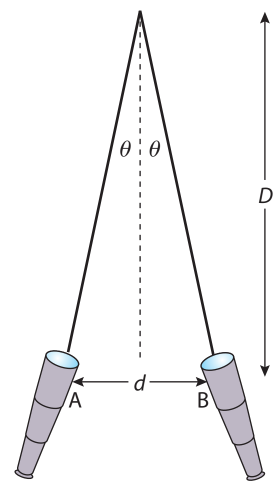


Figure 10.18 A star is viewed at a different angle from two positions. A very distant star would be in the same place.

Stellar parallax

The distance to a point object can be found by measuring the angle between rays from the object to two fixed points. For example, in Figure 10.18, the distance to the star can be found by measuring the angle the telescope is rotated through when it is moved from A to B. The distance D is very large so this angle will be very small; to make it measurable the distance d must be as big as possible, so rather than moving the telescope on the Earth, the movement of the Earth itself is used. The telescope is lined up with the star on one day then again half a year later when the Earth is on the other side of the Sun. To measure the angle, there must be a reference line; in this case we use stars that are so distant that light from them is parallel. This is shown in Figure 10.19.

Since the angle subtended at the star, 2θ is small

$$2\theta = \frac{2\text{AU}}{D} \text{ (in radians)}$$

$$D = \frac{1\text{AU}}{\theta}$$

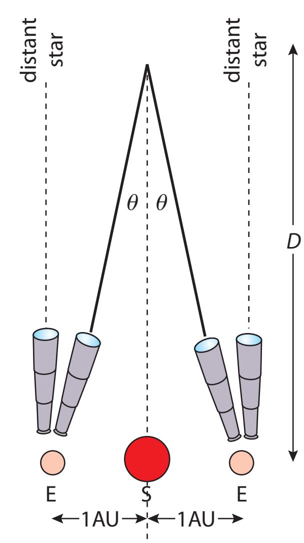


Figure 10.19

Second

One degree can be split up into 60 arc minutes and each arc minute into 60 arc seconds, so there are 3600 arc seconds in one degree.

In radians 1 arc second is therefore $\left(\frac{1}{3600}\right) \times \left(\frac{2\pi}{360}\right)$

To help understanding of stellar parallax, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 10.6.

The parsec

The parsec is defined in terms of the angle subtended at the star.

If the distance to a star is 1 pc, the angle subtended, θ will be 1 second. This makes calculations much easier since the distance in pc is simply $\frac{1}{\text{the angle in seconds}}$

Worked example

The smallest angle that can be measured between light rays that arrive at the surface of the Earth is 0.01 arc second. What is the distance to a star that subtends this angle?

Solution

$$D = \frac{1}{0.01} = 100\,\text{pc}$$

Exercises

12 If the angle subtended by a star is 1 arc sec what is the distance to the star in metres?

13 Calculate the angle subtended by a star that is 5 pc from the Earth.

14 A star subtends an angle of 0.15 arc sec to the Earth. Calculate its distance.

Spectroscopic parallax

Assessment statements

E.3.9 State that the luminosity of a star may be estimated from its spectrum.

E.3.10 Explain how stellar distance may be determined using apparent brightness and luminosity.

E.3.11 State that the method of spectroscopic parallax is limited to measuring stellar distances less than about 10 Mpc.

E.3.12 Solve problems involving stellar distances, apparent brightness and luminosity.

Due to bending of the light by the atmosphere, it is not possible to measure angles smaller than 0.01 arc sec on the Earth. This means that distances greater than 100 pc cannot be measured. Satellite telescopes are better since they are out of the atmosphere; they can resolve angles as small as 0.001 arc sec, extending the range to 1000 pc. Beyond this, other methods must be used one of which utilizes the HR diagram to find the luminosity of the star.

Finding the luminosity using the HR diagram

By measuring the spectra of a star we can determine the most intense wavelength of light emitted. Then using Wien's law, it is possible to calculate the temperature of the star. Now assuming the star is a main sequence star we can use the HR diagram to find its luminosity.

Wien's law

$$\lambda_{max} = \frac{2.90 \times 10^{-3}}{T}$$

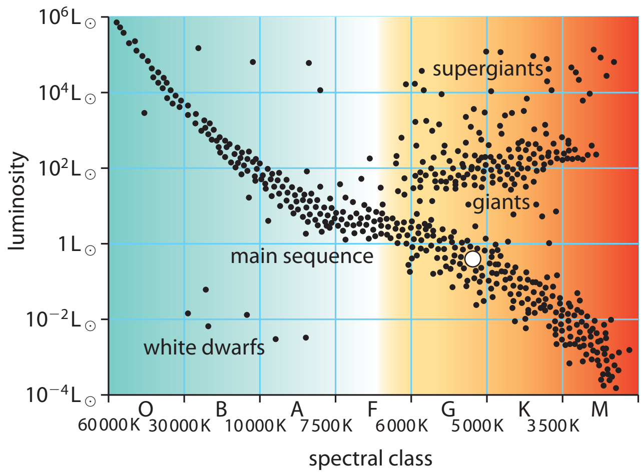


Figure 10.20 The H-R diagram.

If we now measure the apparent brightness of the star we can use the inverse square law to calculate its distance from the Earth.

Worked example

The maximum wavelength of a distant star is measured to be 600 nm and its apparent brightness is $1.0 \times 10^{-12}\,\mathrm{W\,m^{-2}}$. What is its distance from the Earth?

Solution

First we can use Wien's law to find the star's temperature.

$$\lambda_{max} = \frac{2.90 \times 10^{-3}}{T}$$

Rearranging gives $$T = \frac{2.90 \times 10^{-3}}{600 \times 10^{-9}} = 4800\,\mathrm{K}$$

Using the HR diagram we can deduce that if this is a main sequence star, its luminosity is $1\,L_\odot = 3.84 \times 10^{26}\,\mathrm{W}$

The apparent brightness (b) of a star is related to the luminosity by the equation

$$b = \frac{L}{4\pi d^2}$$

where d is the distance from the Earth.

Rearranging this gives

$$d = \sqrt{\frac{L}{4\pi b}} = \sqrt{\frac{3.84 \times 10^{26}}{4\pi \times 1.0 \times 10^{-12}}} = 5.5 \times 10^{18}\mathrm{m}$$

$$= 584\,\mathrm{ly}$$

Exercises

15 The spectrum of a main sequence star has maximum intensity at 400 nm and an apparent brightness of $0.5 \times 10^{-12}\,\mathrm{W\,m^{-2}}$.

(a) Use Wien's law to find the temperature of the star.

(b) Use the HR diagram to find the luminosity of the star.

(c) Calculate the distance from the star to the Earth.

Cepheid variables

Assessment statements

E.3.13 Outline the nature of a Cepheid variable.
E.3.14 State the relationship between period and absolute magnitude for Cepheid variables.
E.3.15 Explain how Cepheid variables may be used as 'standard candles'.
E.3.16 Determine the distance to a Cepheid variable using the luminosity–period relationship.

For stars further than 10 Mpc the amount of light received is too small to accurately determine their temperature. In this case an alternative method is used that involves measuring the time period of a flashing star.

A Cepheid variable is an unstable star that undergoes periodic expansions and contractions, leading to a periodic change in the apparent brightness of the star, as viewed from Earth. This can be represented graphically, as shown in Figure 10.21.

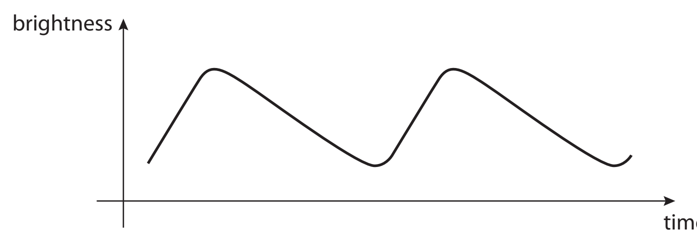


Figure 10.21 The variation of brightness with time for a Cepheid variable. Note how the increase in brightness is faster than the decrease. The star is brightest when it is biggest.

Standard candle
If you know the luminosity of a candle you can use it to measure distance. The further away it is the dimmer it becomes. Since the luminosity of Cepheid variables is known they can be used in the same way so are termed 'standard candles'.

There are many Cepheid variables close enough to the Earth for us to use the stellar parallax method to find their distance. If we then measure their apparent brightness (b) it is possible to calculate their luminosity (L) using the equation

$$b = \frac{L}{4\pi d^2}$$

If the luminosity and period are plotted on a graph, we find that they are directly related as shown in Figure 10.22.

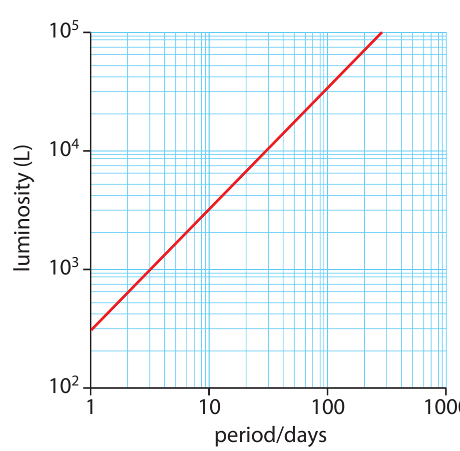


Figure 10.22 The luminosity–period relationship for a Cepheid variable. Note the logarithmic scales.

This is very useful because it means that if we know the star's period, we can use the graph to find its luminosity. Once we know the luminosity, we use the equation above to find its distance from the Earth.

Worked example

The period of a Cepheid variable is 10 days and its brightness $1.0 \times 10^{-10}\,\mathrm{W\,m^{-2}}$. How far is it from the Earth?

Solution

Using the graph in Figure 10.22, the luminosity of a Cepheid variable with period 10 days is $3000L_{\odot}$. The luminosity of the Sun is 3.839×10^{26} W, so this star has luminosity $3000 \times 3.84 \times 10^{26} = 1.15 \times 10^{30}$ W

Rearranging $$b = \frac{L}{4\pi d^2}$$

gives $$d = \sqrt{\frac{L}{4\pi b}} = \sqrt{\frac{1.15 \times 10^{30}}{4\pi \times 1.0 \times 10^{-10}}} = 3.0 \times 10^{19}\,\mathrm{m}$$

$$= 3200\,\mathrm{ly}$$

Exercises

16 A Cepheid variable has period 20 days and brightness $8.0 \times 10^{-10}\,\mathrm{W\,m^{-2}}$. Calculate

(a) its luminosity

(b) its distance from the Earth.

10.6 Magnitude

Assessment statements

E.3.5 Describe the apparent magnitude scale.

E.3.6 Define *absolute magnitude*.

E.3.7 Solve problems involving apparent magnitude, absolute magnitude and distance.

E.3.8 Solve problems involving apparent brightness and apparent magnitude.

The idea of classifying stars according to their brightness originated in ancient Greece two thousand years ago. Since they didn't have any measuring devices the scale was based on how bright the stars appeared to the naked eye. The brightest star was given magnitude +1 and the dimmest +6. In the original scale, a magnitude +1 was twice as bright as *a* +2 and *a* +2 twice as bright as *a* +3 etc. So the difference between *a* +1 and *a* +6 was $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{32}$. However they didn't have any way of measuring the brightness, so it was not very accurate.

Logarithmic scales
A logarithmic scale is one in which each division represents a multiple of some value rather than the sum. For example in Figure 10.22, each division is a multiple of 10 (10, 100, 1000). The magnitude scale is also logarithmic but this time the scale rises in multiples of 2.512.

Apparent magnitude (*m*)

The modern scale uses the same 1 to 6 designation for the stars visible with the naked eye, but is adjusted to be more in line with the measured values. When measured it is found that the brightest stars are about 100 times brighter than the dimmest (not 32 as in the Greek scale). So the scale was set so that *a* +1 was 100 times brighter than *a* +6. Starting from the dimmest each step in magnitude is therefore 2.512 times brighter than the previous (not twice as in the Greek scale). So the difference between a +1 and +6 is $(\frac{1}{2.512})^5 = \frac{1}{100}$.

The apparent magnitudes of a variety of stars and planets is given in the table.

Object	Apparent magnitude
Sun	−26.8
Full Moon	−12.5
Venus (brightest)	−4.4
Venus (dimmest)	−2.7
Sirius	−1.47
Vega	0.04
Betelgeuse	0.41
Polaris	1.99
Pluto	15.1

Limits
The dimmest star you can see with the naked eye is $m = 6$.
With a pair of binoculars $m = 10$.
A large telescope $m = 20$.
A large telescope used to take a photograph $m = 25$.

Magnitude O
The brightest star is Sirius with $m = -1.47$. Vega has $m \approx 0$ with an apparent brightness of $2.32 \times 10^{-8}\,\mathrm{W\,m^{-2}}$.

From this list we can work out how much brighter different stars are than each other.

We have seen that if the apparent brightness of a star is b_1 and its apparent magnitude is m_1 then it will be $\frac{1}{2.512^{m_1}}$ times less bright than a star of apparent magnitude 0 and brightness b_0. In other words:

$$\frac{b_1}{b_0} = \frac{1}{2.512^{m_1}} = 2.512^{-m_1}$$

Likewise for a second star of brightness b_2 and apparent magnitude m_2

$$\frac{b_2}{b_0} = 2.512^{-m_2}$$

So the ratio of the brightnesses

$$\frac{b_1}{b_2} = \frac{2.512^{-m_1}}{2.512^{-m_2}}$$

$$\frac{b_1}{b_2} = 2.512^{m_2 - m_1}$$

● **Hint:** Using logs
To calculate $2.512^{1.58}$ enter 2.512 into your calculator, find the log then multiply by 1.58. The answer is 10^x of this number.

Worked example

Polaris has apparent magnitude 1.99 and Betelgeuse 0.41. How much brighter is Betelgeuse than Polaris?

Solution

$$\frac{b_{\text{Betelgeuse}}}{b_{\text{Polaris}}} = 2.512^{1.99-0.41} = 2.512^{1.58} = 4.3$$

So Betelgeuse is 4.3 times brighter then Polaris.

Exercises

17 Use the values in the table above to calculate how many times brighter Sirius is than Vega.

18 If the brightness of a star of apparent magnitude 0 is $2.52 \times 10^{-8}\,\mathrm{W\,m^{-2}}$, what is the brightness of Polaris?

19 Calculate the apparent magnitude of a star with brightness $1 \times 10^{-10}\,\mathrm{W\,m^{-2}}$ if the brightness of a star of magnitude 0 is $2.52 \times 10^{-8}\,\mathrm{W\,m^{-2}}$

20 How much brighter is an $m = 1$ star than an $m = 25$?

Brightness and luminosity
The apparent magnitude of a star is a measure of its brightness as viewed from the Earth.

The absolute magnitude is a measure of luminosity.

Absolute magnitude (*M*)

The apparent magnitude of a star depends on how far it is from the Earth. The absolute magnitude is the magnitude of a star viewed from a distance of 10 pc, this is more useful when comparing stars.

Since absolute magnitude is directly related to the star's luminosity, the HR diagram can be plotted with M on the vertical axis, instead of luminosity.

If M is the magnitude of a star at 10 pc and m is its magnitude from the Earth then

$$\frac{b(\text{on Earth})}{b(10\text{ pc})} = 2.512^{M-m}$$

But we know that the brightness is proportional to $\dfrac{1}{\text{distance}^2}$

so $$\frac{b(\text{on Earth})}{b(10\text{ pc})} = \frac{10^2}{d^2} = \left(\frac{10}{d}\right)^2$$

therefore $$\left(\frac{10}{d}\right)^2 = 2.512^{M-m}$$

Taking logs: $$2\log\left(\frac{10}{d}\right) = (M - m)\log 2.512$$

Rearranging gives

$$M = m - 5\log\left(\frac{d}{10}\right)$$

Or if we make d the subject

$$d = 10 \times 10^{\left(\frac{m-M}{5}\right)}\text{ pc}$$

So given the absolute and apparent magnitudes of stars we can calculate their distance from the Earth.

Star	Apparent magnitude (m)	Absolute magnitude (M)
Sun	−26.8	+4.8
Sirius	−1.47	+1.4
Vega	+0.04	+0.5
Betelgeuse	+0.41	−5.14
Polaris	+1.99	−3.6

Worked example

Sirius has an apparent magnitude of −1.47 and an absolute magnitude of +1.4. Calculate the distance between Sirius and the Earth.

Solution

Using the formula

$$d = 10 \times 10^{\left(\frac{m - M}{5}\right)}\ \text{pc}$$

$$d = 10 \times 10^{\left(\frac{-1.47 - 1.4}{5}\right)}\ \text{pc}$$

$$\boldsymbol{d = 2.7\ \text{pc}}$$

Exercise

21 Use the table of *M* and *m* to calculate the distance from the Earth to

(a) Sun
(b) Vega
(c) Betelgeuse
(d) Polaris

10.7 Stellar processes and stellar evolution

Assessment statements

E.5.1	Describe the conditions that initiate fusion in a star.
E.5.2	State the effect of a star's mass on the end product of nuclear fusion.
E.5.3	Outline the changes that take place in nucleosynthesis when a star leaves the main sequence and becomes a red giant.
E.5.4	Apply the mass–luminosity relation.
E.5.5	Explain how the Chandrasekhar and Oppenheimer–Volkoff limits are used to predict the fate of stars of different masses.
E.5.6	Compare the fate of a red giant and a red supergiant.
E.5.7	Draw evolutionary paths of stars on an HR diagram.
E.5.8	Outline the characteristics of pulsars.

The birth of a star

Stars are formed when huge clouds of gas and dust are compressed. They can't form on their own because the gravitational force is not big enough to pull the particles together. However, if something causes the cloud to be compressed, such as an exploding supernova or a collision between two dust clouds, the particles get closer and the gravitational force becomes sufficient to start pulling the particles together.

As the gas atoms are pulled together by the gravitational force, they gain KE, and the temperature of the gas increases. This increase in temperature causes an outward pressure that pushes against the gravitational attraction. However, as the atoms get closer, the gravitational attraction increases $\left(F = \frac{Gmm}{r^2}\right)$ so the gas continues to collapse and get hot at an ever-increasing rate.

Fusion starts

As the cloud collapses, a dense core is formed surrounded by a cloud of gas and dust. The centre of the dense core rapidly contracts, resulting in high temperature and pressure. This star, called a *protostar*, gives out light due to its high temperature, but isn't visible because it is surrounded by a cloud of gas.

After about 10^5 years of mass increase, the radiation from the protostar blows away the dust cloud and the mass of the star stabilizes. The star is now a *pre-main sequence star*. The core continues to contract and heat up until the atoms are moving fast enough for fusion to take place. Since hydrogen is so abundant in the universe, it follows that this gas is mainly hydrogen, so the fusion that takes place is the fusion of hydrogen nuclei as shown in Figure 10.7. This proton – proton chain can be summarised in the following equation

$$4{}^1_1\text{H} \rightarrow {}^4_2\text{He} + 2\text{e}^+ + 2\nu_e + 2\gamma$$

The Horsehead Nebula is a vast cloud of gas and dust visible only because it blocks out light from the background glow.

Once fusion starts, the increase in temperature causes greater pressure, balancing the inward force of gravity. The star now stops contracting and becomes a *main sequence* star like the Sun.

The process of fusing hydrogen to helium in the core takes much longer than any other process in the star's life; this is why there are more main sequence stars than any other type. As the star develops, its luminosity and temperature change. This means its position on the HR diagram changes. Figure 10.23 shows this variation for three stars of different mass; these lines are called evolutionary paths.

Since the times involved in the evolution of a star are so great, we cannot observe a star changing from one stage to the next – we can only see stars that are in different phases.

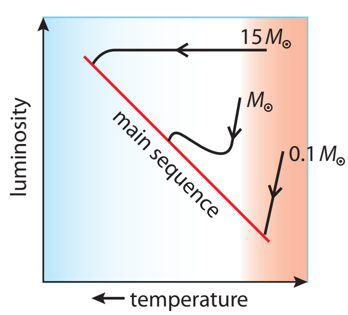


Figure 10.23 The HR diagram for three stars as they turn from protostar to main sequence. Note how the luminosity of the big ones stays constant; this is because they are getting smaller but hotter. A sun size star also shrinks but its outside stays cool until the inside gets so hot that it heats the outer layers. The core of a small star never gets that hot, so it gets less and less bright as it contracts.

Mass-luminosity relationship

The luminosity of massive main sequence stars is greater than stars of small mass; this enables us to know where the different stars join the main sequence line. The equation relating mass, m and luminosity, L is

$L \propto m^{3.5}$

As this process takes place, helium builds up in the core. When all the hydrogen in the core has become helium, the core is no longer producing energy, so the thermal pressure drops, causing the core to rapidly compress. This compression results in an increased temperature, which heats the hydrogen gas outside the core, causing it to start fusing. As this happens the outer layers expand and the star becomes a *giant*.

To find out more about stellar evolution, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on weblink 10.7.

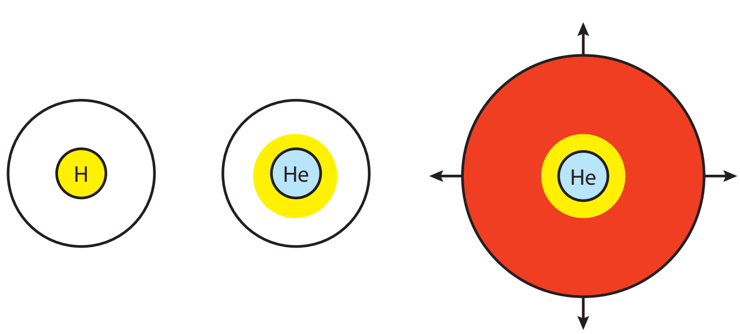


Figure 10.24 The formation of a red giant. The hydrogen core fuses to helium; this then contracts, heating the shell of hydrogen around it, which then fuses, heating the outer layers of gas. As they expand, they cool, giving the star its red colour.

In the meantime more helium is being added to the core, which continues to contract and heat up until the helium starts to fuse into carbon. In the core of a very large star this process continues through many stages of different nuclei fusing until iron is formed. This is the end of the line, since as you remember from the BE curve, iron has the highest $\frac{\text{BE}}{\text{nucleon}}$ ratio so fusing iron would not release energy.

The fate of a star

Small star

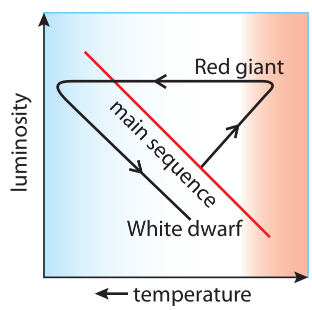


Figure 10.25 The evolution line path for a small star.

Only stars with large mass will continue to fuse nuclei until iron is formed. Smaller stars (up to about 4 $M_\odot$) do not develop the pressure necessary to fuse carbon, so when they have used up all the helium, the core continues to contract, radiating energy as it shrinks. If the star continued to contract, it would disappear altogether. This doesn't happen because the electrons contained in the core cannot occupy the same space at the same time (this is called *electron degeneracy* and is due to the Pauli exclusion principle mentioned in Chapter 15). By the time this point has been reached, the radiation from the core has blown away the outer layers of the star, exposing the core. This now cools down, becoming a white dwarf. The mass of a white dwarf can be no more than 1.4 $M_\odot$. This is called the *Chandrasekhar limit*, and it means that white dwarfs cannot be formed by main sequence stars bigger than 4 $M_\odot$ since they will have cores that are too big.

Big star

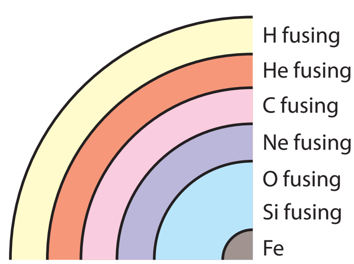


Figure 10.26 The different fusing layers in the core of a big star.

Stars bigger than 4 $M_\odot$ have core pressures and temperatures high enough for carbon fusion to take place. This leads the formation of neon and magnesium. As these heavier elements are formed, they form layers, with the heaviest near the centre as shown in Figure 10.26.

Eventually the larger nuclei fuse to form iron, which forms the central core. Since iron has the highest binding energy, it will not fuse to form larger nuclei, so fusion in the core stops. This means that there is no energy production to give the nuclei enough KE to oppose the crushing pressure caused by gravitational attraction to the outer layers. The core is therefore compressed until electron degeneracy prevents further contraction. There is however a limit to the amount of pressure the electrons can withstand; as more iron is added to the core, its size exceeds the Chandrasekhar limit and electrons begin to combine with protons to form neutrons (like beta decay in reverse) resulting in the emission of a large number of neutrinos. As the core collapses, it reaches the point where it contains only neutrons packed together as close as they would be in a nucleus. The unsupported outer layers now

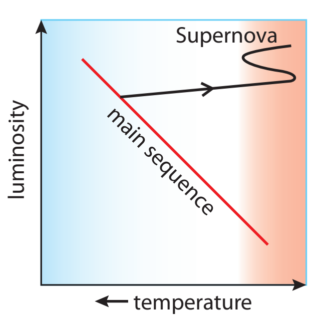
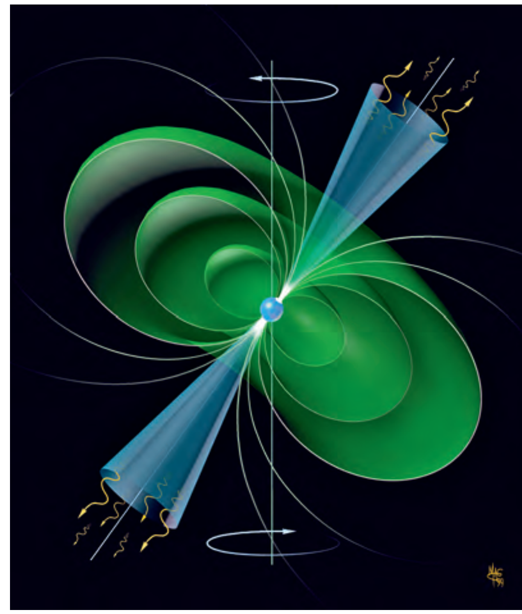


Figure 10.27

An optical image of a supernova 1987a. It's the bright star on the right of the photograph.

come rushing in, bounce off the core and fly back out again in a massive explosion. The star has become a *supernova*. After the explosion all that remains is the neutron core. The star is now a *neutron star*, which will remain stable as long as its mass is not more than 3 $M_{\odot}$ – this is the *Oppenheimer–Volkoff* limit.

The magnetic field of the neutron star sweeps around creating beams of EM radiation along the poles, as shown in this computer illustration.



Pulsars

Most known stars rotate (the Sun has a period of 1 month). This is because there was some rotation in the cloud of dust that formed them. As the cloud is compressed, this rotation is conserved, due to a rotational version of the conservation of momentum. As a star collapses, the speed of rotation gets faster and faster, in the same way as the rotation of an ice skater or ballet dancer gets faster when they pull their arms into their body. A neutron star therefore rotates very fast. Another thing that increases when the star collapses is the magnetic field; if you imagine squashing the Earth into a tiny ball, the magnetic field lines would be very close. As the magnetic field is swept around, it causes charged particles to accelerate, which emit EM radiation (from radio to X-ray) in a fine beam along the line of the poles (this is rather difficult to visualize but it follows Fleming's left hand rule). Since the magnetic poles are not lined up with the axis of rotation, the beam of radiation sweeps around like the beam from a lighthouse. If viewed from Earth, it seems to flash at regular intervals (from 0.03 s to 1.3 s).

Black holes

If the mass of a neutron star is bigger than 3 $M_{\odot}$, neutron degeneracy cannot stop it from collapsing further. In this case it keeps contracting until the gravitational force is so big that not even light can escape; this is a black hole.

The first pulsar was detected by Jocelyn Bell in 1967. It was a radio frequency pulsar with a period of 1.34s. So unusual was this observation, that it was suggested it might be a signal from extraterrestrials

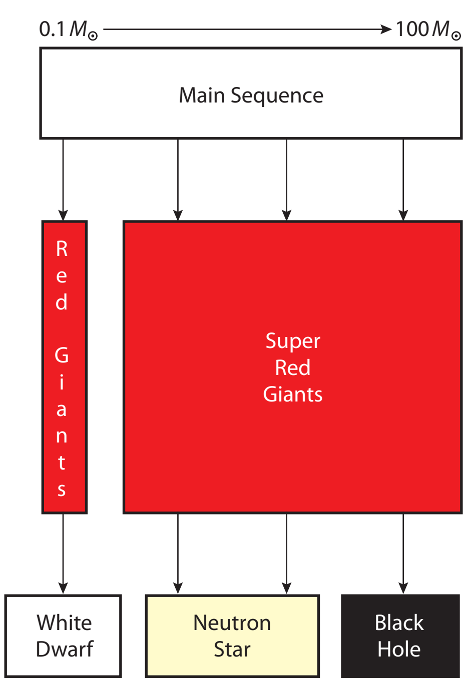


Figure 10.28 The fate of different size stars.

10.8 Cosmology

Assessment statements

E.4.1 Describe Newton's model of the universe.
E.4.2 Explain Olbers' paradox.
E.4.3 Suggest that the red-shift of light from galaxies indicates that the universe is expanding
E.4.4 Describe both space and time as originating with the Big Bang.

E.6.1 Describe the distribution of galaxies in the universe.
E.6.2 Explain the red-shift of light from distant galaxies.
E.6.3 Solve problems involving red-shift and the recession speed of galaxies.
E.6.4 State Hubble's law.
E.6.5 Discuss the limitations of Hubble's law.
E.6.6 Explain how the Hubble constant may be determined.
E.6.7 Explain how the Hubble constant may be used to estimate the age of the universe.
E.6.8 Solve problems involving Hubble's law.
E.6.9 Explain how the expansion of the universe made possible the formation of light nuclei and atoms.

E.4.5 Describe the discovery of cosmic microwave bac kground (cmB) radiation by Penzias and Wilson.
E.4.6 Explain how cosmic radiation in the microwave region is consistent with the Big Bang model.
E.4.7 Suggest how the Big Bang model provides a resolution to Olbers' paradox.

Newton's model

Newton believed that the universe was infinitely big. This would imply that the gravitational force on each star was the same in each direction, holding them in static equilibrium. If the universe is static, then the stars will be in the same place forever and must have been there forever. He also concluded that the universe must be uniform, since if it were not, the forces on every star would not be balanced and there would be movement on some stars as they were pulled together. So Newton's universe was infinitely big, infinitely old, static and uniform.

Olbers' paradox

The problem with Newton's model of the universe was that if the universe were infinite, then when you look into the sky there must be an infinite number of shining stars, and if that is the case, why is the sky dark? You may think the answer is obvious; the light from the distant stars is blocked by the ones in front or some dark stuff that we can't see. However, it's not that simple; as you know, when radiation is absorbed, the energy in it is absorbed, and this would then be re-emitted, adding to the light coming from the blocking stars. The most probable solution to this paradox is that the universe is not infinite, so there aren't an infinite number of stars. Also, time is not infinite and in that case the light from some of them hasn't even reached us yet.

Part of the Virgo cluster of galaxies.

Distribution of galaxies
Galaxies are not distributed randomly about the universe but are found in **clusters**. These galactic clusters are also clustered into **superclusters**.

Red shift

Another nail in the coffin of Newton's model was the discovery that the light from distant galaxies is red-shifted. We have discussed how the shift of lines in the spectrum of a star can be used to deduce whether it is moving towards the Earth or away. A shift towards red implies the star is receding, so if light from galaxies is red-shifted then they must also be receding. It was found that not only is the light from all galaxies red-shifted but the ones that are furthest away are shifted by more than the closer ones. This implies that the universe is expanding. It is as if there was a big explosion, the outer parts flew off fastest and are still travelling outwards with the greatest speed and we are somewhere in the middle. If we look outwards we see the fast galaxies moving away, if we look inwards we see that we are moving away from the slower galaxies, and if we look to the side we see the other galaxies with the same speed moving away as they spread out from the centre. Figure 10.29 illustrates this.

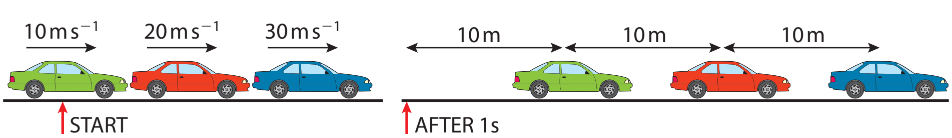


Figure 10.29 As the cars travel at different speeds, the middle one moves away from both the one in front and the one behind (the length of the car has been ignored).

The Big Bang

When presented with the idea that the universe started with a Big Bang, two questions spring to mind – what was there before the Big Bang, and, if the universe is expanding, what is it expanding into, in other words what is beyond the universe?

As we discussed in the first chapter, length and time are quantities we use to define objects and events in our universe. If there is no universe, then there is no length or time. You couldn't fly past in a spaceship to observe the Big Bang, since there was no space to fly through. You can't even say 'when the Big Bang happened', because there was no time to measure the event. Time and space grew out of the

Big Bang. After that moment, space and time were created and expanded, but they didn't expand into some void, since there was no void to fill. So when we say the universe is expanding, we mean that space is growing, rather than it is spreading into the nothingness that surrounds it. The galaxies therefore move apart from each other because the space between them is expanding (like the cars in Figure 10.30) rather than because they are moving with different speeds. This explanation can also be used to explain red shift; if the space through which the wave travels is stretched then the wavelength will increase.

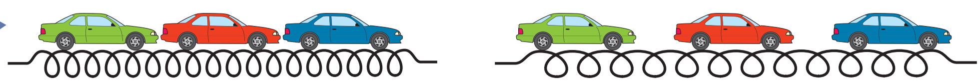


Figure 10.30 The cars move apart because the road is stretching not because they are moving along the road.

Calculating red shift

The velocity of a galaxy relative to Earth can be calculated using the formula

$$\frac{\Delta\lambda}{\lambda} = \frac{v}{c}$$

where $\Delta\lambda$ = change in wavelength
λ = original wavelength
v = relative velocity
c = speed of light

Exercises

22 A spectral line from a distant galaxy of wavelength 434.0 nm is red-shifted to 479.8 nm. Calculate the recession speed of the galaxy

23 The same line from a second galaxy is shifted to 481.0 nm. Calculate its recession speed. Is this galaxy closer or further away?

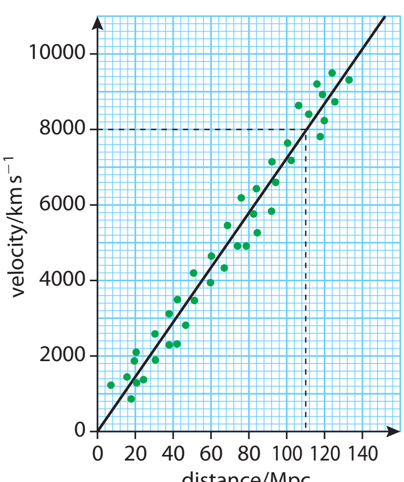


Figure 10. 31 Graph of recession speed of galaxies against their distance from the Earth.

Uncertainties
As you can see in Figure 10.31, the data points are scattered about the best fit line indicating that there are quite big random errors in the measurements.

Hubble's law

We know how to measure the distance between the Earth and distant galaxies and their recession speed. In the 1920s Edwin Hubble plotted these quantities against each other and realized that the recession speed was directly related to the distance to the galaxy. This supports the theory that the universe is expanding. If you consider the cars in Figure 10.29, in one second the blue car has moved 20 m away from the green one, so has a recession velocity of 20 m s^{-1}, whereas the closer red one has only moved 10 m away. The graph in Figure 10.31 shows this relationship.

The gradient of this line is called the *Hubble constant, H_0*. The value of this is not certain but is in the region of 72 km s^{-1} Mpc^{-1}.

$$H_0 = \frac{\text{recessional velocity}}{\text{separation distance}}$$

Exercises

24 Use Hubble's law to estimate the distance from the Earth to a galaxy with a recessional velocity of 150 km s^{-1}.

25 If a galaxy is 20 Mpc from Earth, how fast will it be receding?

The age of the universe

At the time of the big bang all parts of the universe were in the same place, so if we know how fast any two parts are moving apart and how far apart they are now, we can calculate the age of the universe.

$$\text{Age of universe} = \frac{\text{separation distance}}{\text{recessional velocity}}$$

This is the same as $\frac{1}{H_0}$

So the age of the universe $= \frac{1}{H_0}$

To calculate this in seconds we first need to convert the distance into km.

$$H_0 = \frac{72}{3.09 \times 10^{19}} = 2.33 \times 10^{-18}\,\text{s}^{-1}$$

So $\frac{1}{H_0} = 4.29 \times 10^{17}\,\text{s}$

Now converting this into years

$$\frac{1}{H_0} = \frac{4.29 \times 10^{17}\,\text{s}}{3.16 \times 10^{7}} = 1.36 \times 10^{10}\text{ years}$$

This calculation assumes that the velocity is constant. However, we know that gravitational attraction will slow the galaxies down; the recessional velocity we measure today is therefore smaller than it was. This makes our value too large, so according to these measurements, the universe can't be older than 1.36×10^{10} years.

The history of the universe

Just after the Big Bang, the universe was a very small random mixture of fundamental particles and photons, with a temperature of about 10^{32} K. To understand what happened in the next 300 000 years, you need to study the particle physics option. The universe was nothing like it is today, no stars or galaxies, not even atoms. There were no atoms because the temperature was so high that the photons had enough energy to ionize atoms, so any electron combining with a proton to form a hydrogen atom would be knocked off. This event would also be very likely, since at that time the photons were so dense. As the universe expanded, it cooled down until, after 300 000 years, it reached a temperature of 4000 K. This is equivalent to a particle energy of 0.4 eV, not enough energy to ionize hydrogen. At this point electrons started to combine with protons to form atoms. If the universe had spread out uniformly the gravitational attraction between each particle would have been balanced and the particles would not have been brought together to form galaxies; therefore there must have been some irregularities to provide the starting point for all the galaxies that began to form when the universe was 10^9 years old.

The Big Bang is one of several theories of how the universe began. Any competing theories must also explain why the universe is expanding and where the CMB came from.

Cosmic microwave background radiation, CMB

In the beginning, the universe was packed tightly with photons and other fundamental particles. The density of photons was so high during the first 300 000 years that this was called the *radiation-dominated universe*. These photons are still around today, but due to the expansion of space they have a much longer wavelength than before (*cosmological redshift*) – they are in fact now microwaves. This radiation was detected by Penzias and Wilson in the 1960s. At first this was thought to be uniform but the COBE satellite detected very small variations that were just enough to show that the early universe was not completely uniform, thereby enabling galaxies to form.

10.9 The development of the universe

Assessment statements

E.4.8 Distinguish between the terms open, flat and closed when used to describe the development of the universe.

E.4.9 Define the term *critical density* by reference to a flat model of the development of the universe.

E.4.10 Discuss how the density of the universe determines the development of the universe.

E.4.11 Discuss problems associated with determining the density of the universe.

E.4.12 State that current scientific evidence suggests that the universe is open.

E.4.13 Discuss an example of the international nature of recent astrophysics research.

E.4.14 Evaluate arguments related to investing significant resources into researching the nature of the universe.

We have described how the universe was created from the time of the Big Bang, but what will happen from now on? Will it continue expanding, or will it stop?

Open, closed or flat

Open universe: keeps expanding.

Flat universe: the rate of expansion tends to zero at infinite time.

Closed universe: stops expanding and starts to contract.

In the previous section, the age of the universe was estimated by assuming that it was expanding at a uniform rate. Due to the force of gravity we know that cannot be true; the rate of expansion must be slowing down, but whether it is slowing down enough to stop depends on the density of matter within it. If the density is bigger than some critical value(*the critical density*) it will stop expanding and start to contract, eventually ending in the 'big crunch'. If the density is lower than that value, it will continue to expand forever. You may think that a third possibility is that the universe stops expanding and becomes static. However, this is not possible unless the parts are an infinite distance apart, otherwise gravity would always pull them back together; in other words it would keep expanding. These three possibilities are illustrated in the graph of Figure 10.32.

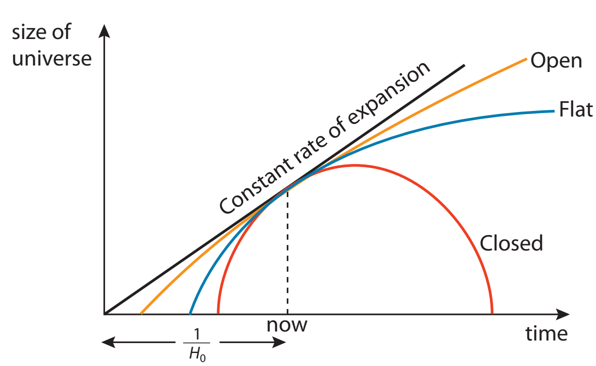


Figure 10.32 Three different possible scenarios depending on the density of the universe. The black line shows the age of the universe as predicted from the Hubble constant.

Critical density
The critical density is the density at which the closed universe becomes open. It is very unlikely that this is the way the universe is, since just one extra electron would make it contract.

Calculating critical density

We can calculate the critical density in a similar way to the way we calculated escape velocity. If we consider a mass m on the edge of the expanding universe, then, as it moves outwards, its KE will be converted to PE. The mass will continue moving outwards until its KE is zero. If this happens at infinity then the PE will also be zero as illustrated in Figure 10.33.

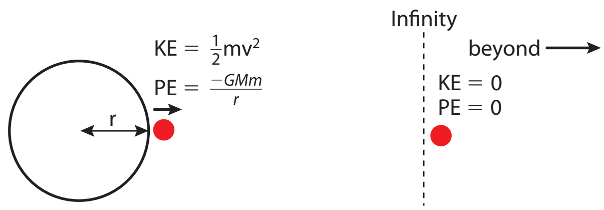


Figure 10.33 Change in KE and PE as the universe expands.

According to the law of conservation of energy
loss of KE = gain in PE

$$\tfrac{1}{2}mv^2 - 0 = 0 - \frac{-GMm}{r}$$

$$\tfrac{1}{2}mv^2 = \frac{GMm}{r}$$

If the critical density of the universe is ρ_0 then the mass M is given by

$$M = \tfrac{4}{3}\pi r^3 \rho_0$$

From Hubble's law we know that $v = H_0 r$

If we substitute these values into the energy equation we get

$$\rho_0 = \frac{3H_0^{\ 2}}{8\pi G}$$

This has a value of about 10^{-26} kg m^{-3} which is equivalent to 6 hydrogen atoms per cubic metre.

Determining the density of the universe

The density of the universe can be measured by measuring the mass of all the stars in a given volume. However, if all the stars and gas clouds in a galaxy are measured, the total mass is not big enough to give the gravitational attraction to hold it together. In fact it's only about 4%. The rest of this mass is called *dark matter*.

Is it possible to make a theory about something you know nothing about?

Dark matter

Dark matter sounds rather sinister but it's just the term for any matter that does not interact with light. This could consist of neutrinos (if they have any mass) or maybe some new particles (WIMPS or MACHOs). Just when we thought we understood the universe we find that there is 96% that we don't know about!

Some of the dark matter may be contained in Massive Astronomical Compact Halo Objects (MACHOs) or Weakly Interacting Massive Particles (WIMPS).

Dark energy

By measuring the acceleration of distant galaxies it has been found that the rate of expansion of the universe is increasing. This has been explained in terms of dark energy which fills all space and causes an outward pressure counteracting the inward force of gravity. This implies that the universe is open as shown in Figure 10.34.

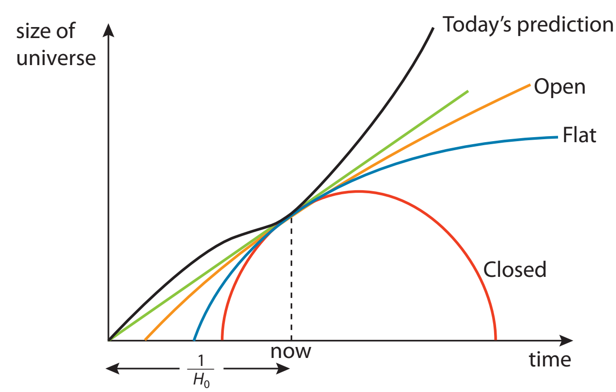


Figure 10.34 The fate of the universe based on an increasing acceleration.

The international nature of astrophysics

In the past we used to refer to the space race. Who would be first into orbit? Who would be first to the Moon? What are the benefits of cooperation rather than competition?

To link to the NASA website, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 10.8.

The study of the universe requires measurements to be taken from as many places as possible; unfortunately we are limited to the Earth and its near surroundings. We would be greatly helped if someone/thing living somewhere else in the universe could send us measurements and observations from where they live. This hasn't happened yet, so for the time being we are limited to the Earth. Given the nature of this research, not to mention the cost, it would be extremely limiting to restrict one's research to the borders of the country in which you live; for this reason, astrophysical projects tend to be international ventures and the data collected is made freely available on the internet. To make it possible to cross reference data, international protocol has to be established so that the naming of objects and measurements is consistent. You can find astrophysical data on many internet sites such as NASA's site which, although an agency of the United States of America, is involved with many international projects, such as the Gamma-ray Large Area Space Telescope (GLAST), which involves the U.S. Department of Energy and institutions in France, Germany, Japan, Italy and Sweden.

Is it worth it?

There are two ways of looking at the question as to whether physicists should be doing research into the origins and ends of the universe. In terms of 'is it worth it for the advancement of science?' the answer is clearly yes, if we are to understand our universe we have to do experiments to gather data. In terms of 'is it worth all the money that is spent?' then the answer is a lot more complex. Whether the money invested in the research will ever be paid back by the discovery of new sources of energy, for example, no one knows. However, according to Douglas Adams in his book *The Hitch Hiker's Guide to the Galaxy*, the answer to Life, the Universe, and Everything is already known – it's 42.

Practice questions

1 This question is about the nature of certain stars on the Hertzsprung-Russell diagram and determining stellar distance.

The diagram opposite shows the grid of a Hertzsprung-Russell (H-R) diagram on which the positions of the Sun and four other stars A, B, C and D are shown.

(a) State an alternative labelling of the axes,

(i) *x*-axis (1)

(ii) *y*-axis (1)

(b) Complete the table below.

Star	Type of star
A	
B	
C	
D	

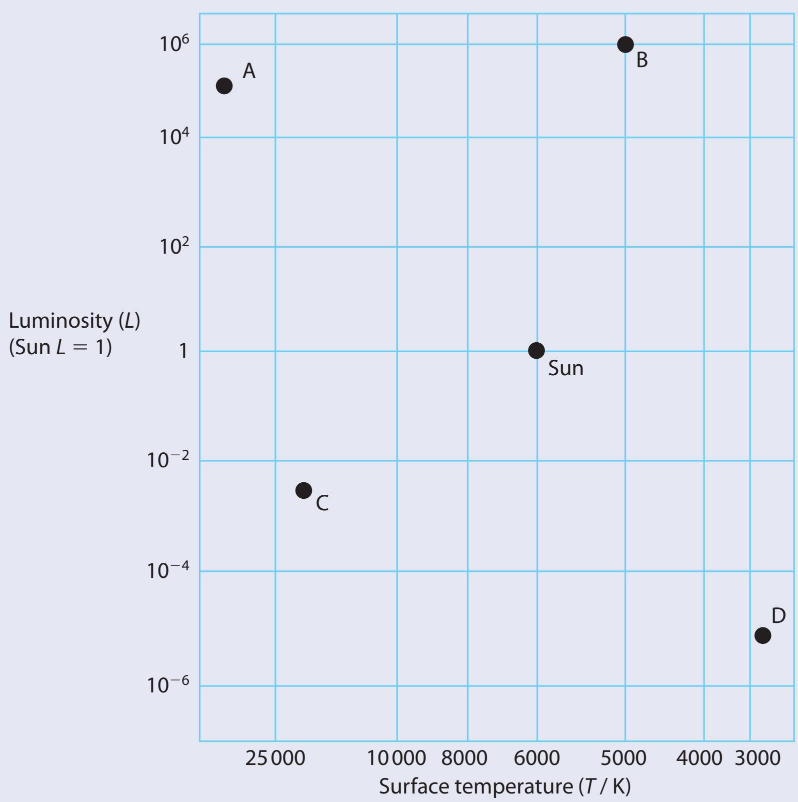
(4)

(c) Explain, using information from the H-R diagram, and without making any calculations, how astronomers can deduce that star **B** is larger than star **A**. (3)

(d) Using the following data and information from the H-R diagram, show that star **B** is at a distance of about 700 pc from Earth.

Apparent visual brightness of the Sun $= 1.4 \times 10^{3}\,\mathrm{W\,m^{-2}}$
Apparent visual brightness of star B $= 7.0 \times 10^{-8}\,\mathrm{W\,m^{-2}}$
Mean distance of the Sun from Earth $= 1.0\,\mathrm{AU}$
1 parsec $= 2.1 \times 10^{5}\,\mathrm{AU}$ (4)

(e) Explain why the distance of star **B** from Earth cannot be determined by the method of stellar parallax. (1)



(Total 14 marks)

2 This question is about some of the properties of Barnard's star.

Barnard's star, in the constellation Ophiuchus, has a *parallax angle* of 0.549 arc-second as measured from Earth.

(a) With the aid of a suitable diagram, explain what is meant by *parallax angle* and outline how it is measured. (6)

(b) Deduce that the distance of Barnard's star from the Sun is 5.94 ly. (2)

(c) The ratio $\dfrac{\text{apparent brightness of Barnard's star}}{\text{apparent brightness of the Sun}}$ is 2.6×10^{-14}.

(i) Define the term *apparent brightness*. (2)

(ii) Determine the value of the ratio

$\dfrac{\text{luminosity of Barnard's star}}{\text{luminosity of the Sun}}$ ($1\,\text{ly} = 6.3 \times 10^4\,\text{AU}$). (4)

(d) The surface temperature of Barnard's star is about 3500 K. Using this information and information about its luminosity, explain why Barnard's star cannot be

(i) a white dwarf. (1)

(ii) a red giant. (1)

(Total 16 marks)

3 This question is about Cepheid variables.

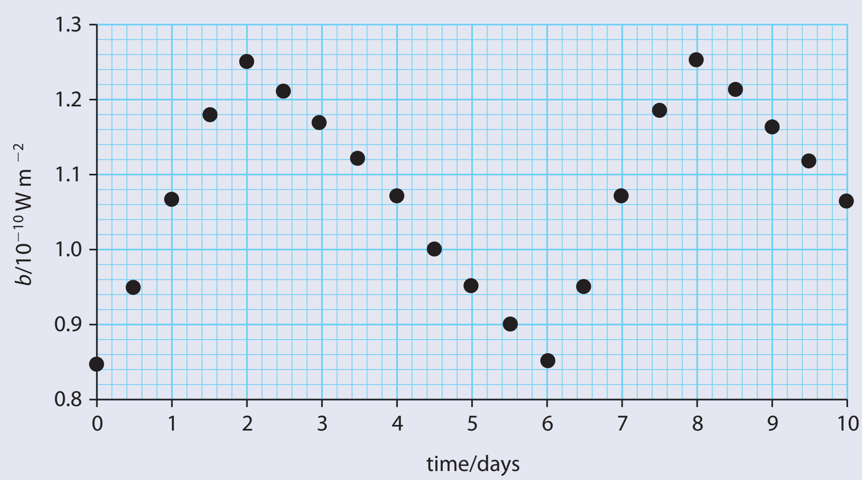
(a) Define

(i) *luminosity*. (1)

(ii) *apparent brightness*. (1)

(b) State the mechanism for the variation in the luminosity of the Cepheid variable. (1)

The variation with time *t*, of the apparent brightness *b*, of a Cepheid variable is shown below.



Two points in the cycle of the star have been marked A and B.

(c) **(i)** Assuming that the surface temperature of the star stays constant, deduce whether the star has a larger radius after two days or after six days. (2)

(ii) Explain the importance of Cepheid variables for estimating distances to galaxies. (3)

(d) **(i)** The maximum luminosity of this Cepheid variable is 7.2×10^{29} W. Use data from the graph to determine the distance of the Cepheid variable. (3)

(ii) Cepheids are sometimes referred to as "standard candles". Explain what is meant by this. (2)

(Total 13 marks)

4 This question is about the properties of the star Arcturus.

The following data is for the star Arcturus.

Distance from Earth / m	Apparent magnitude	Absolute magnitude	Spectral type	Luminosity / W
3.39×10^{17}	–0.1	–0.3	K	3.8×10^{28}

(a) Explain the difference between *apparent magnitude* and *absolute magnitude*. (2)

(b) State and explain, with reference to the data, whether Arcturus would be visible without the aid of a telescope on a clear night. (1)

Techniques for determining stellar distances include the use of stellar parallax, spectroscopic parallax and Cepheid variables.

(c) **(i)** Calculate the distance, in pc, of Arcturus from the Earth. (1)

(ii) State and explain which technique would be most suitable for determining the distance to Arcturus. (2)

(iii) Outline the method you have chosen in your answer to **(c) (ii)**. (4)

(d) State how it may be deduced from the data that the surface temperature of Arcturus is lower than that of the Sun. (2)

The temperature of Arcturus is 4000 K.

(e) Calculate

(i) the surface area of Arcturus. (2)

(ii) the radius of Arcturus. (2)

(iii) the wavelength at which the light from Arcturus has its maximum intensity. (2)

(f) Using your answers to **(e)** deduce the stellar type to which Arcturus belongs. (2)

(Total 20 marks)

5 This question is about the mean density of matter in the universe.

(a) Explain the significance of the *critical density* of matter in the universe with respect to the possible fate of the universe. (3)

The critical density ρ_0 of matter in the universe is given by the expression

$$\rho_0 = \frac{3H_0^2}{8\pi G},$$

where H_0 is the Hubble constant and G is the gravitational constant.

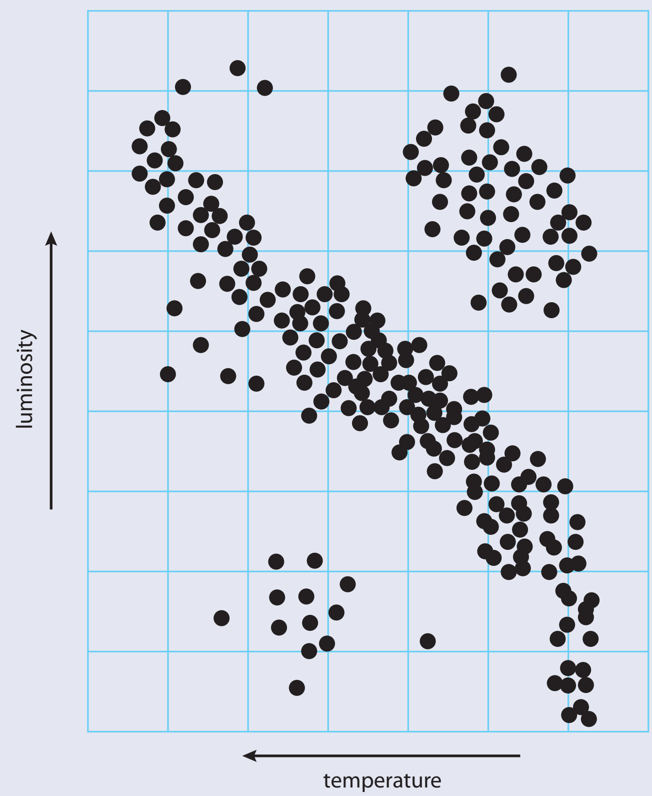
An estimate of H_0 is $2.7 \times 10^{-18}\,\text{s}^{-1}$.

(b) **(i)** Calculate a value for ρ_0. (1)

(ii) Hence determine the equivalent number of nucleons per unit volume at this critical density. (1)

(Total 5 marks)

6 A partially completed Hertzsprung-Russell (H-R) diagram for some stars in the Milky Way galaxy is shown below.



(a) On the diagram,

(i) identify the regions associated with red giants (label the region R) and white dwarfs (label the region W). (1)

(ii) mark with the letter S the approximate present position of the Sun. (1)

(iii) draw the evolutionary path of the Sun from its present position to its ultimate position. (2)

(b) At the end of its main sequence lifetime, a star of approximately ten times the mass of the Sun will start to produce energy at a much higher rate and its surface will become cooler. Outline how it is possible for the star to be producing more power and yet its surface is cooling. (2)

(Total 6 marks)

11 Communications (Option F)

11.1 Radio communication

Assessment statements

- F.1.1 Describe what is meant by the modulation of a wave.
- F.1.2 Distinguish between a carrier wave and a signal wave.
- F.1.3 Describe the nature of amplitude modulation (AM) and frequency modulation (FM).
- F.1.4 Solve problems based on the modulation of the carrier wave in order to determine the frequency and amplitude of the information signal.
- F.1.5 Sketch and analyse graphs of the power spectrum of a carrier wave that is amplitude-modulated by a single frequency signal.
- F.1.6 Define what is meant by sideband frequencies and bandwidth.
- F.1.7 Solve problems involving sideband frequencies and bandwidth.
- F.1.8 Describe the relative advantages and disadvantages of AM and FM for radio transmission and reception.
- F.1.9 Describe, by means of a block diagram, an AM radio receiver.

Communication between people

People communicate by speaking to each other. This is an analogue form of communication since the amplitude and frequency of the voice changes smoothly rather than in steps. Speech can be converted into an electrical signal and transmitted using radio waves.

Radio communication

Radio communication includes not only broadcast radio stations but many other forms of communication that rely on radio signals to transfer information, such as mobile phones, wireless internet and television.

Radio waves

Radio waves are electromagnetic waves that originate from an alternating current. The changing electric field that causes the electrons to move up and down the wire spreads out in all directions. The changing electric field also causes a magnetic field, hence the name electromagnetic wave. The two fields are perpendicular to each other (remember the grip rule).

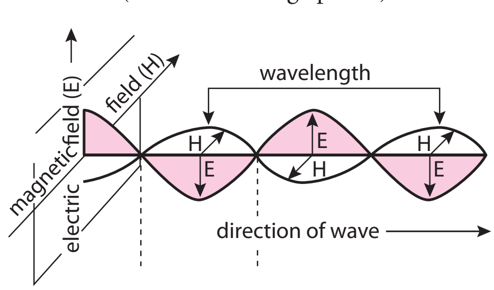
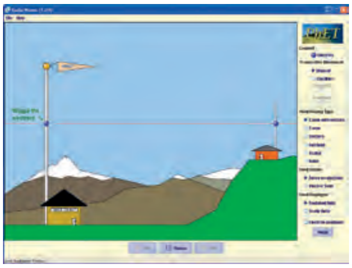


Figure 11.1 A radio wave is an electromagnetic wave.

The phone and the computer use radio communication (but not the cat!).



To see how the electric field spreads out, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 11.1

The radio spectrum

The frequency of the oscillating electrons is the same as the frequency of the wave. This can be anything from a few times a second up to several billion times a second: these frequencies are known as radio frequencies or RF. Different ranges of frequency have different names and uses (see Table 1).

Table 1 The complete radio spectrum.

Name	Frequency	Wavelength	Use
Very low frequency	3–30 kHz	100 km–10 km	avalanche beacons
Low frequency	30–300 kHz	10 km–1 km	Longwave radio
Medium frequency	300 kHz–3 MHz	1 km–100 m	AM radio
High frequency	3–30 MHz	100 m–10 m	Shortwave radio
Very high frequency	30–300 MHz	10 m–1 m	FM radio
Ultra high frequency	300–3000 MHz	1 m–100 mm	TV, Bluetooth, LAN, mobiles

The principle of radio communication

The transmitter

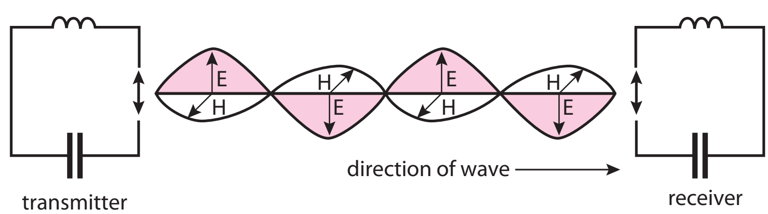


Figure 11.2 A simple radio transmitter and receiver.

A changing electric field must be created in order to produce a radio wave. This can be done using an oscillating circuit, which in its simplest form comprises a coil of wire and a pair of parallel plates (a capacitor). The tuned circuit causes the electrons to oscillate at a single frequency given by the number of turns on the coil and the size of the capacitor, rather like a pendulum whose frequency is fixed by its length. When electrons oscillate in the circuit, a radio wave will radiate in all directions. This circuit is called the *transmitter*.

The receiver

If a second oscillator circuit is placed at a distance from the transmitter it will be in a region of changing electric and magnetic field. Electrons in the circuit will be caused to oscillate due to the changing field. The amplitude of this oscillation is however very small, especially if the transmitter is a long distance away. However, if the frequency of the second oscillator is the same as the first, then resonance occurs, resulting in a much larger amplitude, enough to be detected.

Communication

So we have communication between the two circuits but all we have communicated is a single frequency. A single frequency of sound would be a continuous whistle. When communicating with sound we change the frequency of the sound to make different words, but if we changed the frequency of a radio wave then the receiver wouldn't be able to receive it, since it is tuned in to one specific frequency. The simplest way to communicate would be to repeatedly switch the current on and off. This is how early communication was carried out. Morse Code is a way of representing each letter and number by a series of dots and dashes (a dot is a short pulse and a dash is a long one). In this way text messages can be sent: *dot dot dot*, *dash dash dash*, *dot dot dot* is still an internationally recognized distress signal standing for SOS (save our souls). Changing a wave to carry a signal is called *modulation*. If many people in the same area want to communicate by radio waves then each one needs to choose a different frequency; this is called a *channel*.

There are two common forms of modulation used to broadcast radio programmes, these are *frequency modulation* and *amplitude modulation*; this is what the letters FM and AM stand for on a radio tuner.

Amplitude modulation

Amplitude modulation is when the amplitude of the carrier wave (the radio wave) is varied at the same rate as a signal (see Figure 11.3).

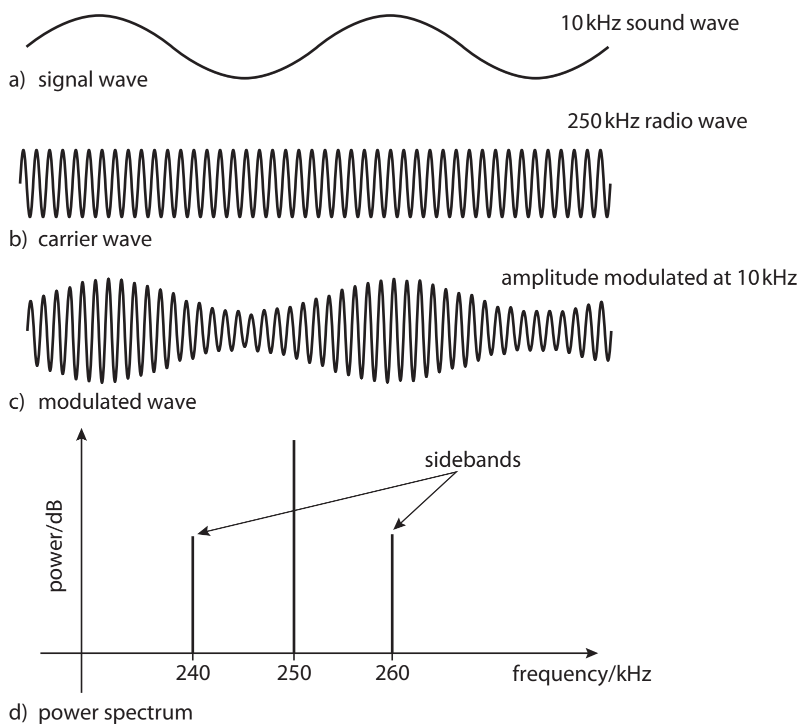


Figure 11.3 A carrier wave modulated by a signal wave. In this example the 250 kHz carrier will have two sidebands 260 kHz (250 + 10) and 240 kHz (250 − 10). The sidebands can be represented on a power spectrum as shown here.

Sidebands

You may remember that when a body oscillates with SHM its displacement can be described by the formula:

$x = x_o \sin \omega_1 t$ where x_o is the amplitude and ω_1 is the angular frequency, $2\pi f$

If x_o also varies sinusoidally then we can write $x_o = A \sin \omega_2 t$

The resultant amplitude is then $x = A \sin \omega_2 t . \sin \omega_1 t$

Now, $\sin A \sin B = \frac{1}{2}(\cos(A-B) - \cos(A+B))$

So $x = \frac{1}{2}A(\cos(\omega_2 t - \omega_1 t) - \cos(\omega_2 t + \omega_1 t))$

In other words, the oscillation is made up of two components: one with angular frequency $(\omega_2 + \omega_1)$ the other with $(\omega_2 - \omega_1)$.

Amplitude modulation
You can try varying the carrier and signal frequencies to see what happens to the side bands. To try this, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 11.2.

A similar thing happens with the wave, modulating the amplitude causes the carrier frequency to split into two components, these are called *sidebands*. The sidebands do not affect the ability of the receiver to resonate with the signal but it does mean that radio channel frequencies cannot be too close to each other.

Power spectrum

A power spectrum (see Figure 11.3d) is a chart showing the relative powers of the different frequencies that make up a signal. This is like the frequency analyser found on many music systems. The power scale is in decibels, which is a logarithmic scale. An increase of 1 bel means that the power has increased by a factor of 10. The central line in the spectrum shown in Figure 11.3d therefore represents a much bigger power than the sidebands.

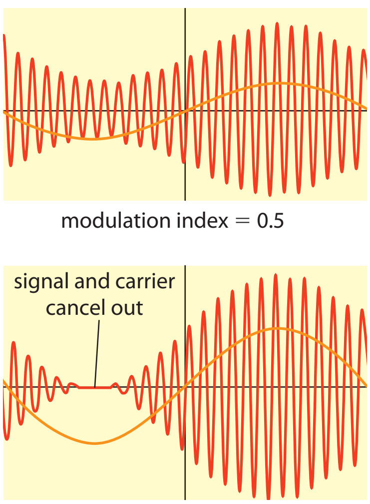


Figure 11.4 With modulation index 1 the signal and carrier wave can completely cancel each other out.

Modulation index

The modulation index is the ratio of the signal amplitude to the carrier amplitude. A high modulation index will cause the final signal to vary greatly.

Bandwidth

In the previous example a single frequency sound of 10 kHz was transmitted. More than just one frequency needs to be used to transmit complex sounds such as music. The complete range of frequencies that humans can hear is from 20 Hz to 20 kHz. Modulating the carrier between these two frequencies would result in a wide band of frequencies from 230 kHz to 280 kHz as in Figure 11.5.

So the signal will occupy a band of frequencies from 230 to 280 kHz. This means that there can't be any other radio channels in this range. It is however possible to reduce the bandwidth by cutting out the highest and lowest frequency notes from the music, but this will result in a loss in quality. Medium wave radio channels are allocated a 9 kHz band width. Telephone uses a narrower bandwidth of 3 kHz since only the human voice is transmitted. An analogue TV signal, on the other hand, has a bandwidth of 6 MHz since information about both pictures and sound needs to be transmitted. Greater bandwidth implies more information can be transmitted.

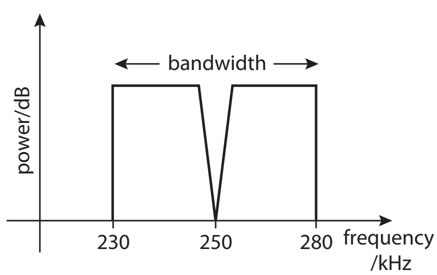


Figure 11.5 The power spectrum for a signal modulated by a band of frequencies.

Broadband internet
The term bandwidth is also used in digital communication to represent the rate of transfer of data. It is similar to the analogue bandwidth in that it is related to the amount of information in the signal but it isn't the same thing. Broadband Internet is therefore a signal that has a lot of information per second.

Exercises

1 If the amplitude of a signal of frequency 500 kHz is modulated by a frequency of 15 kHz, what will the frequency of the sidebands be?

2 Refer to the graph of the modulated signal shown in Figure 11.6.
(a) What is the frequency of the carrier signal?
(b) What is the frequency of the signal wave?
(c) What sidebands will be present in this signal?

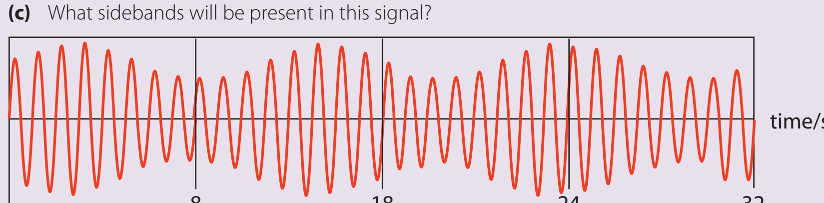


Figure 11.6

Frequency modulation

Frequency modulation also superimposes a signal onto a carrier radio wave but instead of varying the amplitude of the carrier, the frequency is varied between two values either side of the carrier frequency (see Figure 11.7).

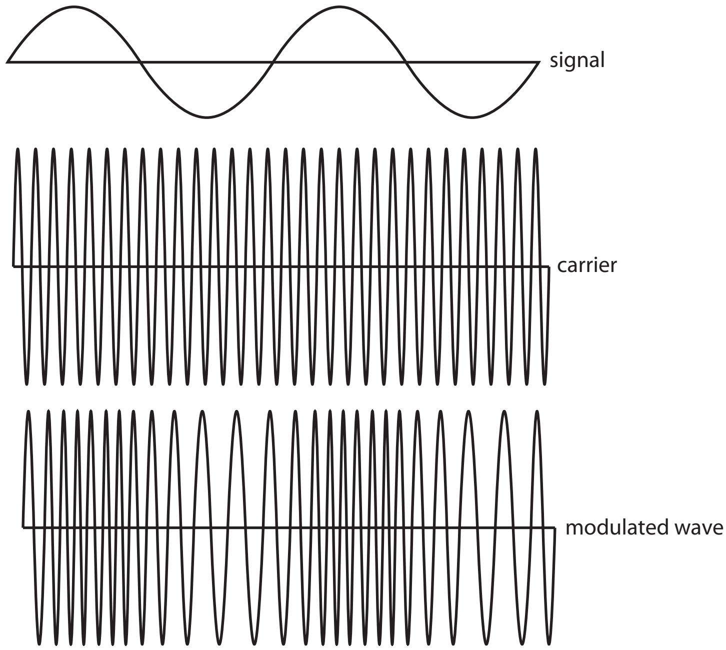


Figure 11.7 In an FM signal the frequency of the carrier, f is modulated by the signal so that its frequency varies continuously between two values, f_1 and f_2.

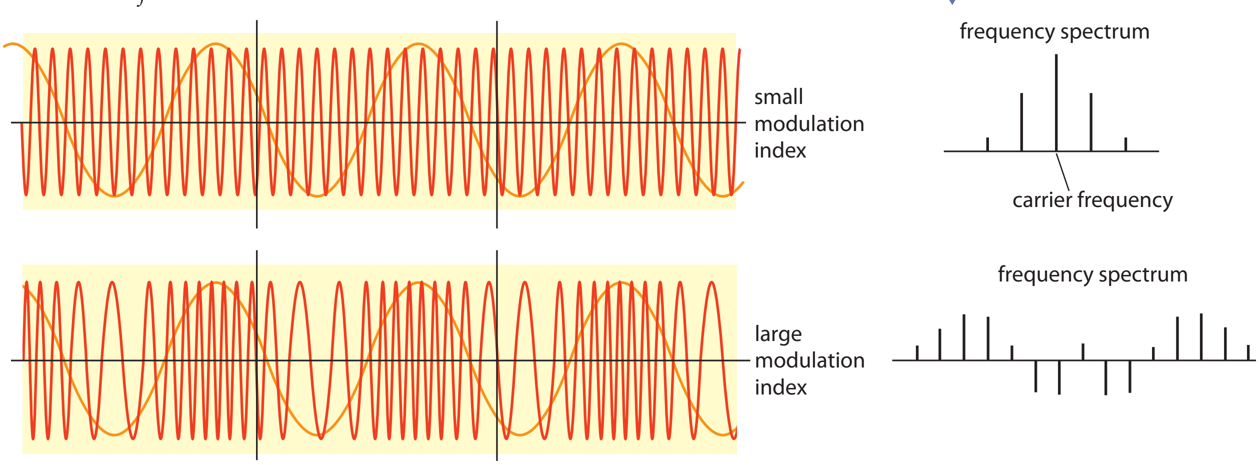
Peak frequency deviation

In an FM signal, the carrier wave deviates from its original frequency f by an amount Δf. This is known as the *peak frequency deviation*. Therefore the maximum frequency $f_2 = f + \Delta f$ and the minimum $f_1 = f - \Delta f$

Modulation index

If the carrier frequency, f is modulated between f_1 and f_2 then the modulation index $= \dfrac{(f_2 - f_1)}{f}$

Figure 11.8 FM signals with a large and a small modulation index and the frequency spectrum and power spectrum respectively.



Sidebands

As in AM, the addition of the modulated wave can be split into components. These components can be displayed on a frequency spectrum (see Figure 11.8).

Bandwidth

As can be seen from these examples, FM signals occupy a bandwidth of frequencies. The width of this band is actually infinite. However, as you can see from the frequency spectrum in Figure 11.8, the power of the bands gets smaller and smaller, until after a certain point they are so small that they can be ignored. The effective bandwidth can be calculated using Carson's Rule.

Bandwidth $= 2 \times$ (peak deviation $+$ highest modulating frequency)

Example

If a carrier signal of 100 MHz is modulated between with a peak deviation of 50 kHz to carry an audio signal with a maximum frequency of 20 kHz then the bandwidth $= 2 \times (50 + 20) = 140$ kHz

Frequency modulation
You can try varying the carrier and signal frequencies to see what happens to the sidebands. To try this, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 11.3

Exercises

3 A 150 MHz signal is modulated between 149.9 MHz and 150.1 MHz to carry a signal ranging from 100 Hz to 20 kHz. Calculate:
(a) the peak frequency deviation
(b) the modulation index
(c) the band width.

4 For the signal in Figure 11.9 calculate:
(a) the highest and lowest modulation frequencies
(b) the signal frequency.

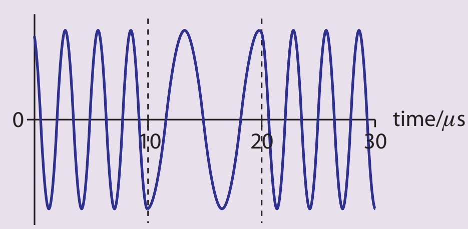


Figure 11.9

Comparing AM and FM

Bandwidth

The bandwidth allocated to AM radio is 9 kHz, just enough to transmit a reasonable quality audio signal. FM uses a bandwidth of 200 kHz. This means that when you tune into FM radio channels they are further apart than AM. This is also why FM uses a higher frequency, although the actual frequencies used by FM and AM are decided by international agreement, not the laws of physics. To use a frequency band, a radio station must pay a sum of money to the controlling body in that country. Two stations can use the same band but they must be geographically far enough apart so that their signals don't interfere.

Range

The range of an AM signal is much longer than an FM signal. This isn't because of the modulation but because the frequencies used by AM are much lower than those used by FM. Low frequency radio waves have two ways of travelling from A to B, either along the ground, or reflected off an atmospheric layer called the *ionosphere*. The latter enables the waves to travel long distances, so two AM radio stations using the same bandwith must be a long way apart. On the other hand, high frequency waves travel in straight lines and don't reflect off the ionosphere. Rather than travelling round the Earth they shoot off into space. This means that to receive FM you have to be able to have direct contact with the transmitter. Satellite TV also uses high frequency radiation, and this is why a receiving dish has to be pointed straight at the satellite: this is called *line of sight*. The good thing about this is that FM stations using the same frequency can be close together, and this is why local radio stations tend to use FM.

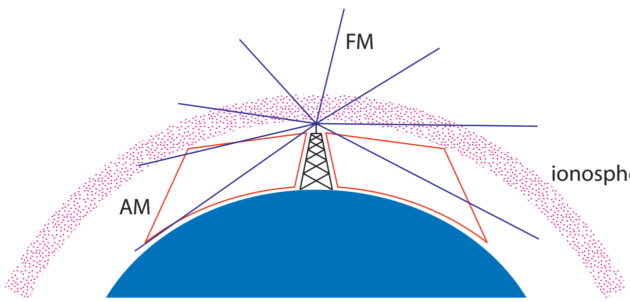


Figure 11.10 AM waves reflect off ionosphere and travel along the ground whereas FM waves travel in straight lines.

Quality

Radio channels are not the only source of radio waves, as any movement of charge will cause EM radiation, e.g. lightning. These signals will interact with the radio signal causing a change in amplitude but not frequency. These disturbances (noise) affect the signal carried by an AM channel but not an FM channel. This is the main problem with AM, it is very susceptible to noise. This is why high quality music channels tend to use FM not AM.

Cost

A simple AM radio receiver can be built for very little money out of a crystal diode, a capacitor and a coil of wire. An FM receiver is rather more complicated and therefore costs more.

The AM radio receiver

In this course you do not have to know how each component in a radio receiver works but you do have to understand the stages. These can be represented in a block diagram as shown in Figure 11.11.

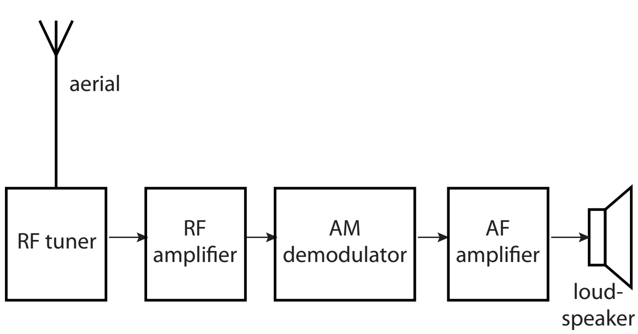


Figure 11.11 The AM radio receiver and its component parts.

Aerial

The aerial is a long conducting rod. Radio waves passing through it cause a changing electric and magnetic field in it.

RF tuner

Anywhere on Earth there will be radio waves from many different sources, each of different frequency and all of which will cause a changing field in the aerial. The radio frequency tuner is manually set to resonate with just one of those frequencies. This is what you do when you tune your radio.

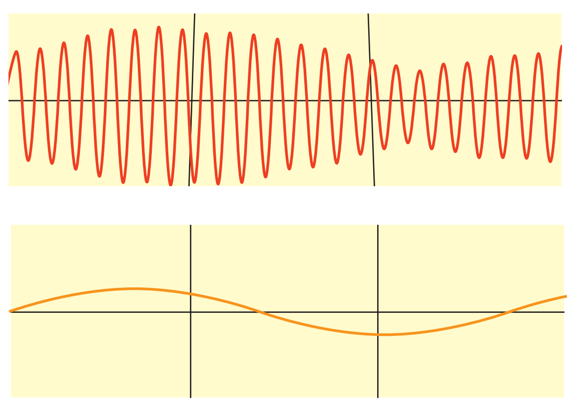
RF amplifier

The signal from the tuner is very small so needs to be amplified before the signal is passed onto the demodulator.

AM demodulator

The demodulator splits the audio signal from the carrier signal. Only the audio signal continues to the next stage, the RF signal is discarded.

Figure 11.12 Signal before and after demodulation.



AF amplifier

This increases the amplitude of the audio frequency signal.

Loudspeaker

This turns the signal into music and/or speech.

11.2 Digital signals

Assessment statements

F.2.1 Solve problems involving the conversion between binary numbers and decimal numbers.

F.2.2 Distinguish between analogue and digital signals.

F.2.3 State the advantages of the digital transmission, as compared to the analogue transmission, of information.

F.2.4 Describe, using block diagrams, the principles of the transmission and reception of digital signals.

F.2.5 Explain the significance of the number of bits and the bit-rate on the production of a transmitted signal.

F.2.6 Describe what is meant by time-division multiplexing.

F.2.7 Solve problems involving analogue-to-digital conversion.

F.2.8 Describe the consequences of digital communication and multiplexing on worldwide communications.

Communication between computers

A computer is a digital device, everything it does is coded from a sequence of 1s and 0s. It is not obvious how 1s and 0s can be used to communicate information, until you understand the binary system.

The binary system

Our normal system of numbers is called the decimal system. In this system we can represent any number by using 10 symbols (0123456789). The way it works is that we group units, tens, hundreds and so on. So 365 is 3 hundreds, 6 tens and 5 units. Units, tens and hundreds are all powers of 10. 10^0, 10^1, 10^2 etc. This is called *base 10* (see Table 2).

Power of 10	3	2	1	0
	1000	100	10	1
	thousands	hundreds	tens	units

Table 2 The decimal system of numbers.

Power of 2	3	2	1	0
	8	4	2	1
	eights	fours	twos	ones

Table 3 The binary system of numbers..

The base of our number system doesn't have to be 10. We can use any number. If we use 2 this is called the binary system, as represented in Table 3. With 4 digits we can count from 0 to 15, with 8 digits we can count to 255 (see Table 4).

Table 4 The numbers 1–10 in binary. In this system the number 7 would be one 4, one 2 and one.

Decimal	Binary
1	00000001
2	00000010
3	00000011
4	00000100
5	00000101
6	00000110
7	00000111
8	00001000
9	00001001
10	00001010

Bit (b)
Each 1 or 0 is called a binary digit or bit .

LSB
This is the least significant bit, this is the digit representing the digit to the right, this represents the smallest part (1s).

MSB
The most significant bit is the digit on the left, this represents the biggest part.

Byte (B)
A byte is 8 binary digits. 8 digits can represent any number from 0 to 255.

Exercises

5 Convert the following numbers into binary form:
(a) 12
(b) 26
(c) 33

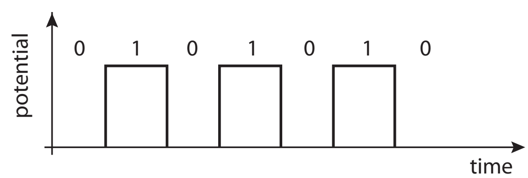
6 Convert the following binary numbers into base 10:
(a) 00110
(b) 11001
(c) 11110

7 What are the most significant bits for the numbers in Question 6?

Sending the data

When computers communicate, binary numbers are sent as a series of high and low electrical potentials (see Figure 11.13). When receiving a series of 1s and 0s it is impossible to know what numbers they represent unless you know how many bits there are per number. So digital devices are standardized by the number of bits they use; this is usually a multiple of 8.

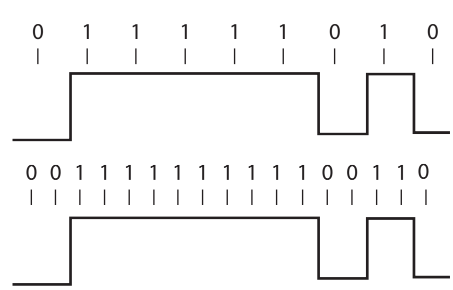
Figure 11.13 Changes in potential represented by 1s and 0s where a high is represented as 1 and a low represented as 0.



Timing

A digital signal is easy enough to read if the sequence is 0101010 as in Figure 11.13. Often it's as complicated as the sequence in Figure 11.14. This could be 011111010 or 00111111111100110 depending on how long each pulse is. A computer uses a very fast clock to read the signal each time it expects a pulse. If information is sent at 20 kb s^{-1} then each pulse lasts 1/20000 s, so the computer needs to know when to start reading and it must read the signal every 50 μs. If the pulses don't come at exactly the right time or if the clock isn't exact then the computer will start to read the wrong information.

Figure 11.14 Two different readouts with two slightly different timings.



Synchronous transmission

In synchronous transmission the sender and receiver clocks are set at the same rate so the transmission of the pulses is regular. A clock signal is sent with the data in order to synchronize the clocks. No clock is 100% accurate, so information is sent in blocks of up to 64 000 bits, so the clock does not have to stay synchronized for too long. There is, however, some room for error since the computer doesn't have to take its reading exactly in the centre of each one or zero.

Asynchronous transmission

If the clocks are not synchronized then the signal must contain information to inform the receiver how often to take readings. There is more chance of error in this system so the block lengths must be much smaller. This system can only be used for slow transfer of data. All fast transfer is synchronous.

Time division multiplexing

Because digital information is sent in blocks, it is a relatively simple process to send different pieces of information in different blocks. In this way many signals can be sent along the same cable. All that has to be added is some code before each block to say which blocks go together and where they should go.

Figure 11.15 Three time division multiplexed signals.

Compression

If the time duration of each pulse is reduced then it is possible to transmit the signal in less time; this is called *compression*.

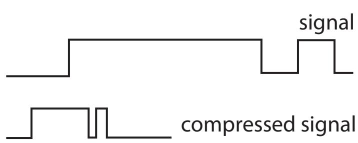
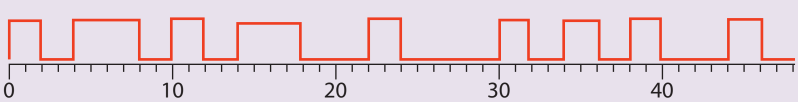


Figure 11.16 A signal and its compressed version. The minimum size of a pulse depends on the speed of the transmitter and receiver.

Exercises



This shows a digital signal. The time scale is in μs.

8 What is the binary data in this signal if it is read at a rate of 500 kHz.

9 If the signal contains two multiplexed signals each 6 μs long, work out the binary data in each of the signals.

10 Convert each of the 3-bit binary numbers in Question 9 to base 10.

MP3

An MP3 file is a compressed file. A piece of music recorded onto a CD takes up 10 times more space that it does when compressed into MP3 format.

Difference between analogue and digital

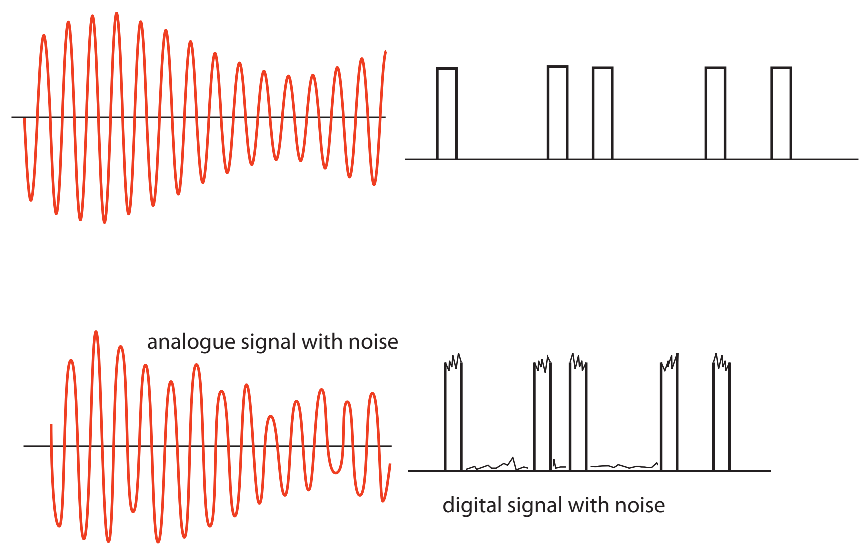
We have considered two different forms of communication, people communicating via radio waves and computers communicating through wires. The signals sent in these examples are fundamentally different; the radio signal carrying music changes continuously from one amplitude to another (AM) or one frequency to another (FM). This continuous variation is called *analogue*. On the other hand, a digital signal is either high or low, there is no variation between the two.

Advantages of digital signals

Noise

When an electrical signal travels from A to B it is affected by changes in electric field that take place between A and B. This could be caused by lightning, electric machinery etc. When an analogue signal is changed by noise, it changes the signal, so music that is affected by noise will sound different. When a digital signal is affected by noise, the information carried is unaltered. Figure 11.17 shows the effect of noise on the signals shown. The analogue signal can be seen to be quite different whereas the digital one is just the same.

Figure 11.17 The effect of noise on an analogue signal and digital signal.



Compression

If a music signal were to be compressed, the pitch of the music would become higher. Compressing a digital signal does not change the information, it just enables it to be transmitted more quickly.

Multiplexing

It is much simpler to multiplex digital signals than analogue ones.

Source independence

A digital signal is always a series of ones and zeros, independent of what type of information is being transmitted. This means speech, text, music and video can all be transmitted on the same signal.

Coding

By changing the sequence of the 1s and 0s in a predetermined way, a digital signal can be coded to prevent anyone not knowing the information (spies) from reading the message.

Data manipulation

A computer is a powerful tool that can only accept digital data. An advantage of using a digital signal is that we can use the computer to manipulate our data.

For these reasons it is very useful to convert analogue signals into data signals. A device designed to do this is called an *analogue to digital converter* (ADC).

The analogue to digital converter (ADC)

When a sound is converted into an electrical signal, a continuously varying p.d. is created. This can be transmitted through a wire and is an analogue signal.

Sampling

To convert an analogue signal into a digital one, it must be changed from a continuously changing p.d. to one that changes in steps (a PAM signal). This is done by measuring the p.d. at regular intervals (sampling) then rounding those p.d.s to the nearest whole number (quantizing).

Table 5 The numbers 1–6 in binary.

Decimal	Binary
1	00000001
2	00000010
3	00000011
4	00000100
5	00000101
6	00000110

By sampling the signal in Figure 11.18 once each second, it can be represented by the following numbers: 0,1,4,5,5,2,0,2,4,1. These can then be converted into binary form. Since the highest number is only 5 these can be represented by just 3 bits. So the first four numbers would therefore be 000 001 100 101.

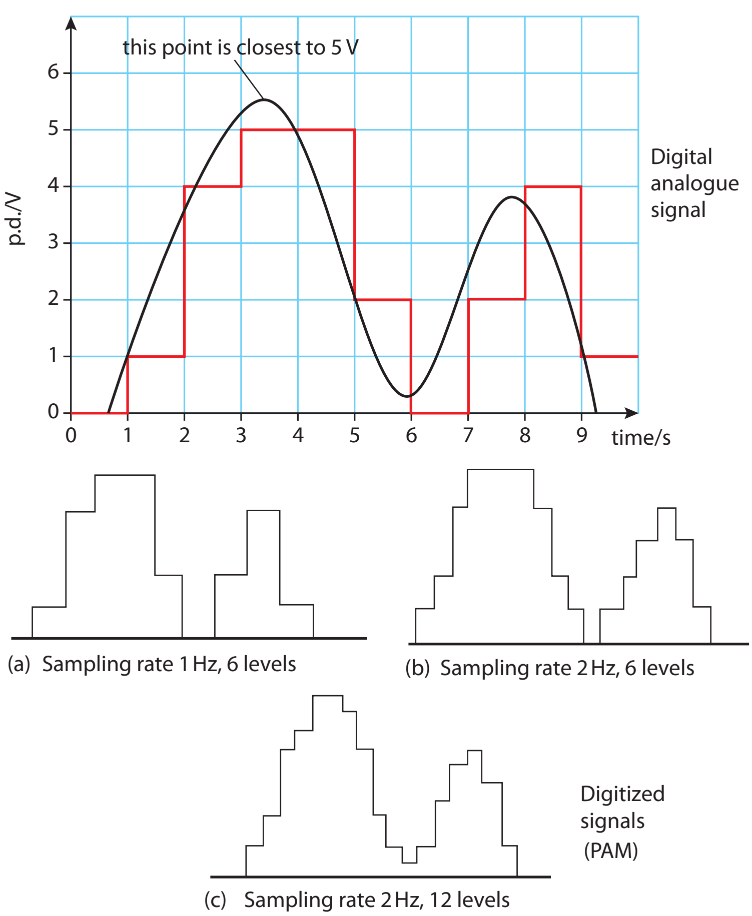


Figure 11.18 Converting an analogue signal into quantized levels.

Sampling rate

When this data is received, the original signal can be reconstructed. This is not exactly the same as the original signal. However it could be improved by increasing the sampling rate as in Figure 11.18b.

Number of bits

It is also possible to reduce the gap between the steps by creating more levels to each volt. The representation of each volt by one digit can be increased to 2 digits per volt giving 12 quantized levels as in Figure 11.18c. Increasing the number of quantized levels means that we need to increase the number of bits to 4. By increasing the sampling rate and using sufficient quantized levels, the signal can be quite close to the original. This system is used in digital sound recording.

Digital music (44.1kHz 16bit)

Digitally recorded music is very close to the original because the signal is sampled 44 100 times each second and the quantized levels are stored as 16-bit binary numbers. With 16 bits you can make 65 536 different numbers and so the equivalent number of quantized levels.

Exercises

11 Using a sampling rate of 1 Hz and 8 quantized levels, convert the analogue data shown in Figure 11.19 to a digital signal (you don't need to turn the signal into binary form, just give the answer as a list of base 10 numbers).

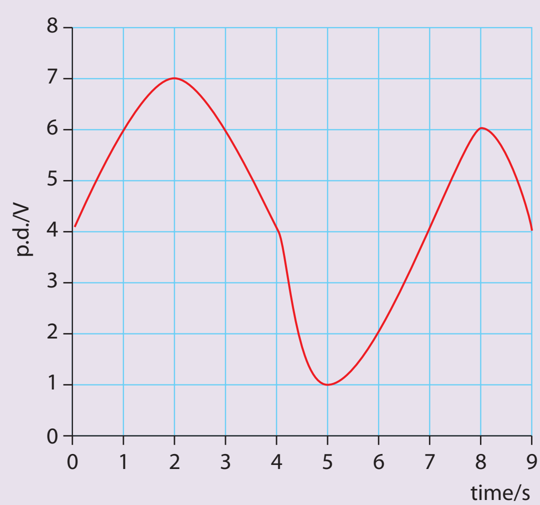


Figure 11.19

12 Convert the first 4 s of data into 4-bit binary.

13 If a 5 minute piece of stereo music is recorded on a 16-bit 44 kHz MP3 player, how many bits of information will be recorded? How many bytes is that?

14 Use your answer to Question 13 to calculate how many hours of music can be recorded with 1 Gbyte of storage. If this is changed to MP3 format and it is compressed by a factor of 10, how many hours of music can you store on a 1 Gbyte MP3 player?

The complete system

The complete digital communication system can be represented by a block diagram.

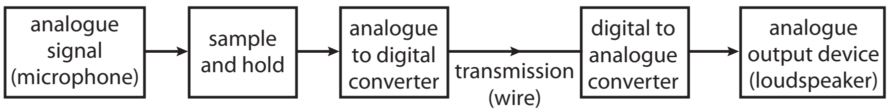


Figure 11.20 A block diagram for a digital communications system and its component parts.

Analogue signal

The original information to be communicated often starts off as an analogue signal, e.g. music. This must first be converted to an electrical signal by a microphone.

Sample and hold

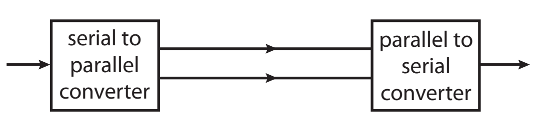
The ADC works by comparing the instantaneous potential with a set number of predetermined quantized levels. This process cannot be performed instantaneously so the signal needs to be held for a short time. The sample and hold stage involves sampling the potential and holding the information so the ADC has time to process it. The time of the hold depends on the sampling rate of the ADC; if it is 44.1 kHz then the signal will be held for 20 μs.

Analogue to digital converter

The ADC takes the steady potential from the sample and hold stage and converts this to the nearest number of quantized levels. This is then converted to a binary number. The number of binary digits depends on how many quantized levels there are, but it is generally a multiple of 8 bits. The clock rate is also included in this data or sent on a separate cable, this is the rate at which the bits are sent. If the sampling is done at 44 kHz and converted to 32 bits then the clock rate would be $32 \times 44\,\text{kHz} = 1.4\,\text{MHz}$.

Transmission

The digital signal can now be transmitted to the receiver. This can be through a single wire, optical fibre or radio wave (more about this in the next section). The signal is said to be serial since all the information is in a line. However it can be sent along parallel wires if first changed into a parallel signal with a serial to parallel converter.



Digital to analogue converter

The DAC decodes the binary information and clock data to convert the series of 1s and 0s to a changing potential. If the receiver is a digital device such as a computer this stage is not needed.

Exercise

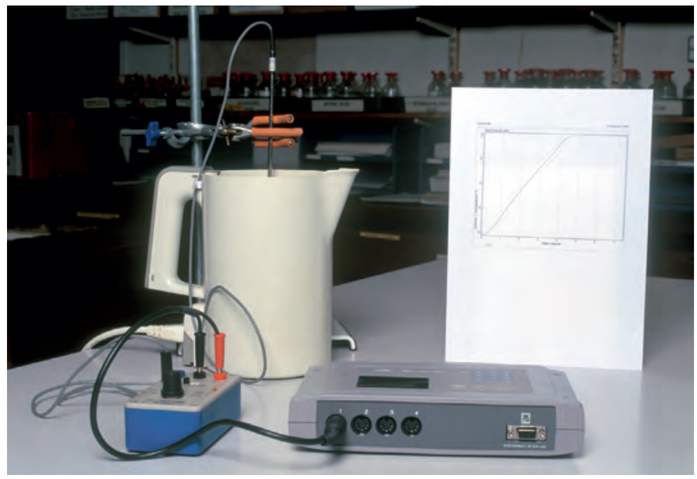
15 If an ADC converts an analogue signal to 16 bits at a rate of 1 MHz

(a) for how long must the signal be held in the sample and hold stage?

(b) what is the clock rate of the ADC?

The use of the interface in physical measurements

It is likely that you have used a computer interface or data logger during the practical part of your physics course. This is a device that takes a signal from a sensor and passes it on to a computer. The information coming in to the interface is always a potential; the interface converts this to a binary code so a computer can process it.



A data logger.

A photo gate is a digital sensor.

Digital sensors

A photo gate is an example of a digital sensor. This is because the photo gate has only two outputs: high or low and nothing in between. The photo gate consists of an IR beam and a detector. When the detector receives the beam, a 5 V potential is created, and if the beam is blocked the potential is 0 V. The interface records the time of a change, converts this into binary code and sends the signal to a computer.

Analogue sensor

Most sensors used in physics experiments are analogue sensors. These produce a continually varying potential. An example of an analogue sensor is a temperature sensor. This gives out a potential that is proportional to the temperature. The interface samples this potential, converts it to a binary code and sends it to the computer.

A temperature sensor is an analogue sensor.

Sampling rate

The sampling rate of the ADC is set by the computer software. A rapidly changing quantity will need to be sampled often; a typical school interface will be able to sample up to 250 000 times a second. If the signal is alternating then the sampling rate must be more than twice the frequency of the signal. You can get some very strange results if you don't sample fast enough, as shown in Figure 11.21.

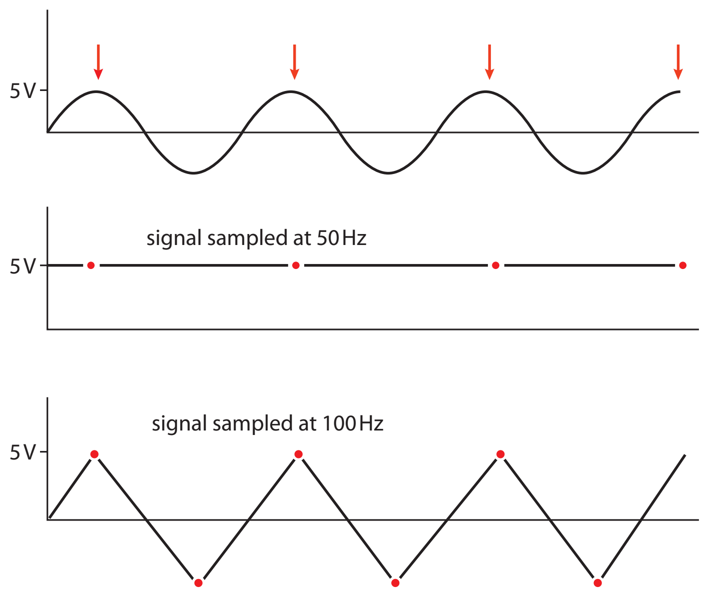


Figure 11.21 A potential alternating at 50 Hz is sampled at 50 Hz and 100 Hz .

Exercises

16 If a data logger can sample at a maximum rate of 10 kHz, what is the highest frequency signal that it can measure?

11.3 Optic fibre transmission

Assessment statements

F.3.1 Explain what is meant by critical angle and total internal reflection.

F.3.2 Solve problems involving refractive index and critical angle.

F.3.3 Apply the concept of total internal reflection to the transmission of light along fibres.

F.3.4 Describe the effects of material dispersion and modal dispersion.

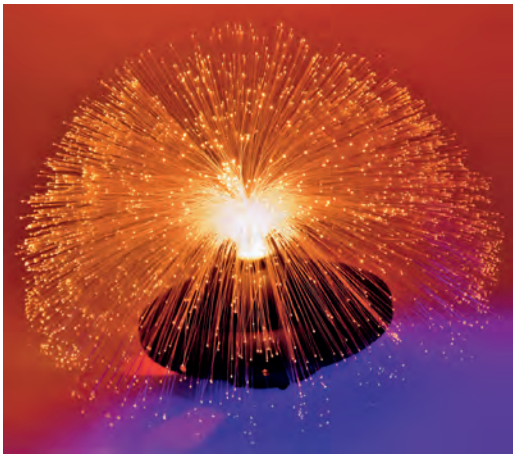
F.3.5 Explain what is meant by attenuation and solve problems involving attenuation measured in decibels (dB).

F.3.6 Describe the variation with wavelength of the attenuation of radiation in the core of a monomode fibre.

F.3.7 State what is meant by noise in an optic fibre.

F.3.8 Describe the role of amplifiers and reshapers in optic fibre transmission.

F.3.9 Solve problems involving optic fibres.

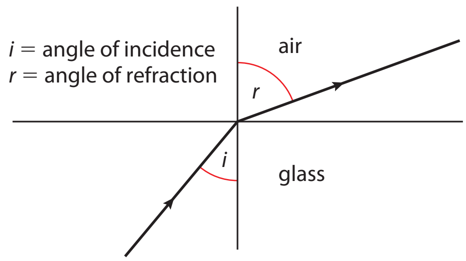


The light emitted from the ends of the fibres make a nice pattern.

Refraction of light

One way of sending a digital signal is by flashing a light on and off (on-1 off-0). To transmit this signal an optical fibre is used. To understand how this works, we first need to know a bit more about refraction of light. Light has different speeds in different materials. If light passes from one material to another, the change in speed causes the light to change direction. If the speed of the light increases, it is deflected towards the boundary between the two materials as in Figure 11.22.

Figure 11.22



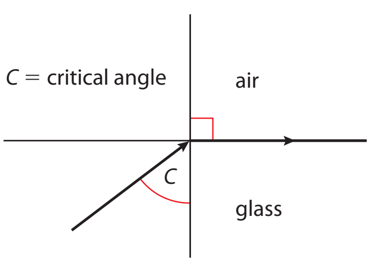
(a) Light refracted from glass to air.

Snell's law tells us that

$$\frac{\sin i}{\sin r} = \text{the refractive index from glass to air}$$

The critical angle

If the angle of incidence increases, a point will be reached where the refracted ray is refracted along the boundary. The angle at which this happens is called the *critical angle*.



(b) Light refracted at 90°.

Applying Snell's law to this situation:

$$\frac{\sin C}{\sin 90} = \text{refractive index from glass to air}$$

$$\sin C = \text{refractive index from glass to air}$$

Refractive index is usually measured from air to glass, but

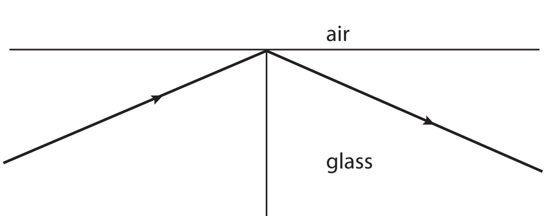
$$\text{Refractive index (air-glass)} = \frac{1}{\text{Refractive index (glass-air)}}$$

$$\text{So } \sin C = \frac{1}{\text{Refractive index (air-glass)}}$$

$$\text{Refractive index(air-glass)} = 1.5$$

$$\text{So } \sin C = \frac{1}{1.5}$$

$$C = 42°$$



(c) Light totally internally reflected.

Total internal reflection

If the critical angle is exceeded, all of the light is reflected. This is known as *total internal reflection*. Since all the light is reflected none is transmitted. This is not the case when light is reflected off a mirror when some is absorbed.

Optical fibre

An optical fibre is a thin strand of glass or clear plastic. If a ray of light enters its end at a small angle, the ray will be total internally reflected when it meets the side. Since the sides are parallel, the ray will be reflected back and forth until it reaches the other end as in Figure 11.23. Optical fibres are used extensively in communication.

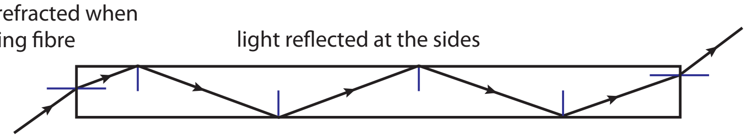


Figure 11.23 Light reflected along a fibre.

Step indexed

Actual fibres have two layers. The light reflects at the boundary between the layers, this is so fibres can be bundled together. This is called *step indexed* since there is a step in refractive index between the core and sheath.

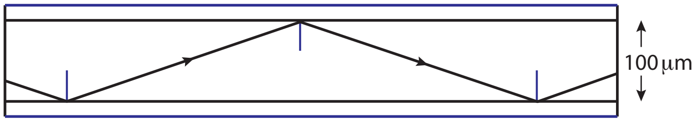


Figure 11.24 A step-indexed multimode fibre.

Modal dispersion

Not all rays of light enter one end of the fibre and come out of the other, only certain waves make it. The possible paths are called *modes*.

The more direct modes reach the end first. This can cause problems with data transfer as bits of data might not arrive in the correct order. It is possible to make very thin fibres that only allow one mode; these are called *monomode fibres*.

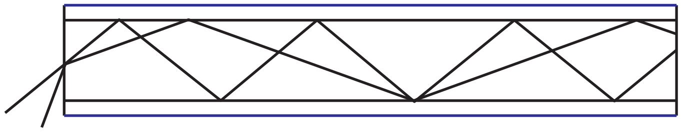


Figure 11.25 There are different possible paths along a fibre.

Material dispersion

When light passes through a prism, different frequencies are refracted by different amounts, causing the colours to disperse. When light travels down a fibre, different frequencies have different paths, as in Figure 11.26.

This will cause a problem if different bits of data arrive at the wrong time.

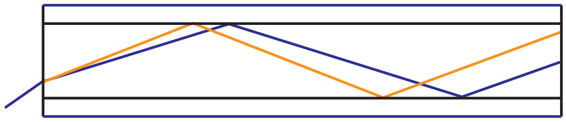


Figure 11.26 Different frequencies have different paths.

Exercises

17 Light enters a glass block of refractive index 1.5 at an angle 70° as shown in Figure 11.27.

(a) Use Snell's law to calculate the angle of refraction θ_1.

(b) Use geometry to find the angle θ_2.

(c) Calculate the critical angle for glass.

(d) Will the ray be totally internally reflected?

(e) Calculate length *D*.

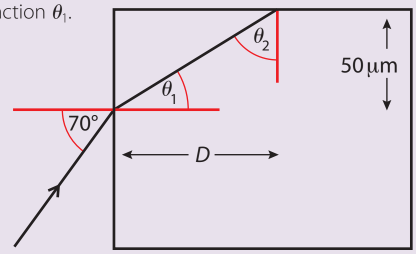


Figure 11.27

Attenuation

Attenuation is the opposite of amplification. If a signal is attenuated its power gets less. Attenuation is measured in decibels (dB).

The decibel

A decibel is a measurement of power with relation to some fixed value. The scale is logarithmic rather than linear; in other words, a change from 1 dB to 2 dB is not the same as a change from 2 dB to 3 dB.

When we use dB to measure attenuation, it gives a measure of how much the power of a signal has been reduced.

The way it is calculated is with the formula:

$$\text{Attenuation} = 10\log_{10}\left(\frac{P_{\text{in}}}{P_{\text{out}}}\right)$$

where P_{in} is the original power and P_{out} the power after attenuation. The factor 10 is to convert from bels to decibels.

Table 6 Remember $\log_{10}x$ is the power you would have to raise 10 to in order to make *x*.

$\text{Log}_{10}(1) = 0$	$10^0 = 1$
$\text{Log}_{10}(10) = 1$	$10^1 = 10$
$\text{Log}_{10}(100) = 2$	$10^2 = 100$
$\text{Log}_{10}(1000) = 3$	$10^3 = 1000$

Table 7 Attenuation table

Power in (mW)	Power out (mW)	Attenuation (dB)
1	0.9	0.46
1	0.8	0.97
1	0.7	1.55
1	0.6	2.22

Worked example

If the power in to a fibre is 1 mW and the power out is 0.01 mW, calculate the attenuation.

Solution

$$\text{Attenuation} = 10\log_{10}\left(\frac{1}{0.01}\right) = \mathbf{20\ dB}$$

Attenuation in fibres

The attenuation of an optical fibre is related to its length, so when calculating the attenuation of a fibre, attenuation per kilometre is calculated.

Fibre type	Attenuation
Multimode	0.8 dB/km
Monomode	0.3 dB/km

Worked example

If the attenuation of a fibre is 0.3 dB km^{-1} what is the power of a 1 mW signal after 5 km?

Solution

If attenuation is 0.3 dB km^{-1} then attenuation in 5 km is 1.5 dB

From the definition of dB, $1.5 = 10\log_{10}\left(\frac{1}{P_{out}}\right)$

$0.15 = \log_{10}\left(\frac{1}{P_{out}}\right)$

$10^{0.15} = \frac{1}{P_{out}}$ $\quad$ 10^x is the inverse of $\log_{10}x$.

$1.41 = \frac{1}{P_{out}}$

$P_{out} = \mathbf{0.71\ mW}$

Attenuation and wavelength

The attenuation of light in an optical fibre is different for different wavelengths. To minimize attenuation, wavelengths with low attenuation should be used.

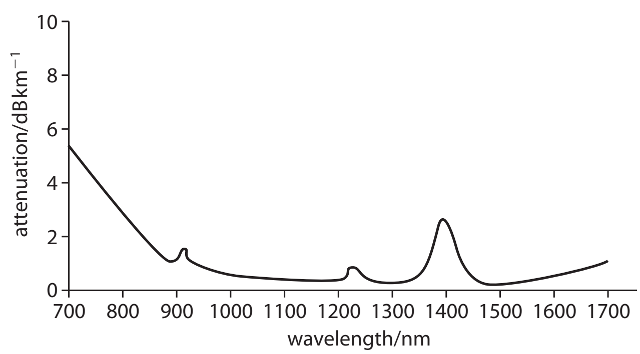


Figure 11.28 This shows attenuation against wavelength for an optical fibre.

Exercises

18 The power into a fibre is 1 mW. Calculate the attenuation if the power out is:
(a) 0.1 mW **(b)** 0.2 mW **(c)** 0.01 mW

19 The attenuation of light in a given fibre is 2 dB km^{-1}.
(a) What is the attenuation after 5 km of fibre?
(b) If a signal of 1 mW is sent into 5 km of fibre, what is the power of the signal that comes out?

20 By looking at the graph in Figure 11.28, decide if a wavelength of 1400 nm would be a good choice for sending a signal along this fibre.

Amplification in optical fibres

As light travels along a fibre its power becomes less. For long-distance communication the signal therefore needs to be amplified. Amplification can be achieved by inserting a length of fibre that has the addition of atoms that give out light when the signal passes. This is called *stimulated emission* and is used in the operation of lasers. The added (doped) atoms are selected so that the light they emit is the same wavelength as the signal. They are first excited (pumped) by a laser, and then when the signal passes they give out light in the same direction and in phase.

This simulation from PhET shows how a laser works. To view, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 11.4.

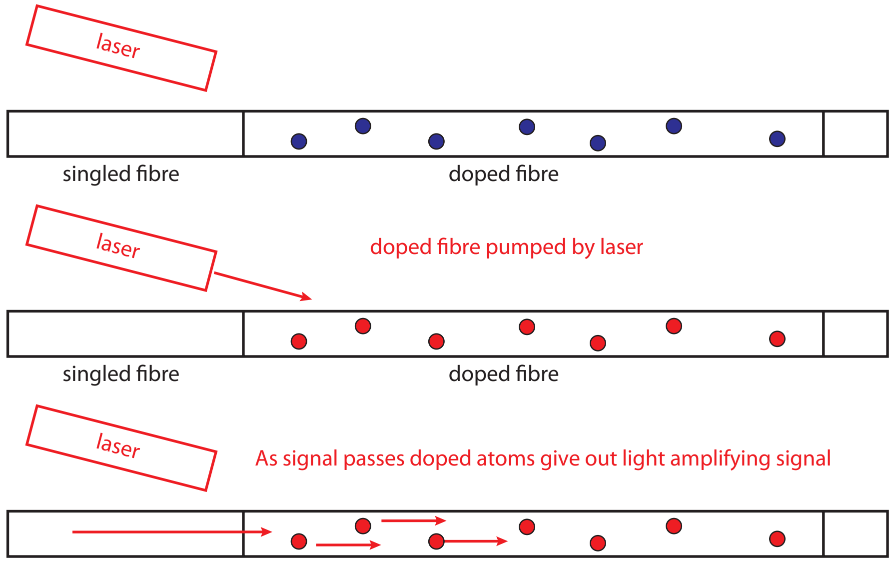


Figure 11.29 A doped fibre used to amplify the signal. The doped atoms are excited by a laser. The signal causes the doped atoms to emit radiation.

Noise

Since light is not a radio frequency it is not affected by interference with RF sources, however there is some noise created on amplification. Some of the atoms pumped can give out light before the signal arrives. The effect of this noise can be reduced by reshaping the signal.

Comparing optical fibres to copper cables

Compared to copper coaxial cables, optical fibres:

- have much greater bandwidth due to higher frequencies used
- have less attenuation
- have greater security since it is very difficult to tap into an optical fibre
- have no noise due to RF interference
- are lighter and thinner
- are more expensive.

11.4 Channels of communication

Assessment statements

F.4.1 Outline different channels of communication, including wire pairs, coaxial cables, optic fibres, radio waves and satellite communication.

F.4.2 Discuss the uses and the relative advantages and disadvantages of wire pairs, coaxial cables, optic fibres and radio waves.

F.4.3 State what is meant by a geostationary satellite.

F.4.4 State the order of magnitude of the frequencies used for communication with geostationary satellites, and explain why the up-link frequency and the down-link frequency are different.

F.4.5 Discuss the relative advantages and disadvantages of the use of geostationary and of polar-orbiting satellites for communication.

F.4.6 Discuss the moral, ethical, economic and environmental issues arising from satellite communication.

F.2.9 Discuss the moral, ethical, economic and environmental issues arising from access to the internet.

A digital signal is a varying potential which can be sent from A to B by connecting a conducting wire between A and B. The wire is called a *communication channel*. Applying a changing potential to one end of the wire causes a changing potential at the other. Alternatively, the signal can be changed into electromagnetic radiation and sent as a radio signal.

Types of cable

Wire pairs

Wire pairs are the cheapest form of cable, consisting of two insulated wires that can either be parallel or twisted. The main problem with this form of cable is that the strength of the signal gets less as the length of the cable increases. This is attenuation. There are several reasons for this:

- Resistance: the resistance of a wire is proportional to its length and causes energy to be lost in the form of heat. This can be minimized by making the wires out of a material with low resistivity.
- Radiation losses: the changing electric field in the wire creates a changing electric and magnetic field that radiates from the wire This can also lead to interference between the wires or 'cross talk'.

Attenuation also increases with frequency so wire pairs can only be used for long range low frequency signals such as telephone or short range high frequency signals as in the communication between computers. Wire pairs can transfer data at a rate of several Mb s^{-1} over short distances.

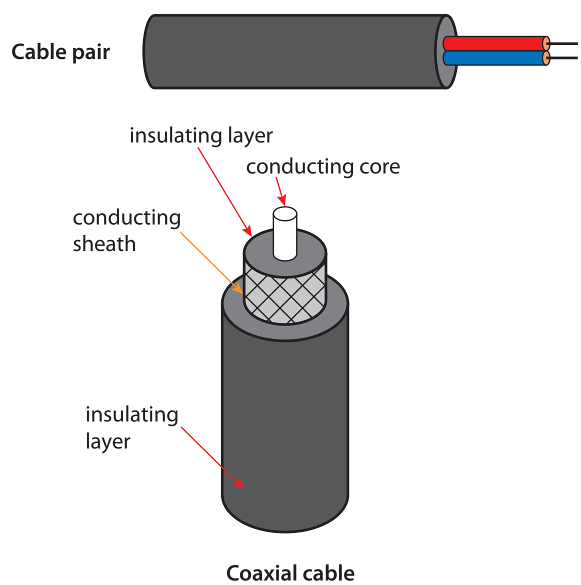


Figure 11.30 Comparison of a cable pair and coaxial cable.

Coaxial cables

The two conductors in coaxial cables consist of a core surrounded by a sheath; the two are separated by an insulator.

The attenuation in a coaxial cable is much less than in a wire pair because the outer sheath prevents the loss of signal by radiation. It also prevents disturbance of the signal from outside sources. Coaxial cables are used for transmission of the TV signal from the aerial to the TV set and also for long range internet connections. It is possible to transfer data at a rate of several Gb s^{-1} over short distances with a coaxial cable.

Using radio waves to transmit digital information

Radio waves are a continuous variation of electric and magnetic fields so they are themselves analogue signals. However, they can be used to transmit a digital signal in the same way as they are used to transmit an analogue signal. This is done by modulating the wave.

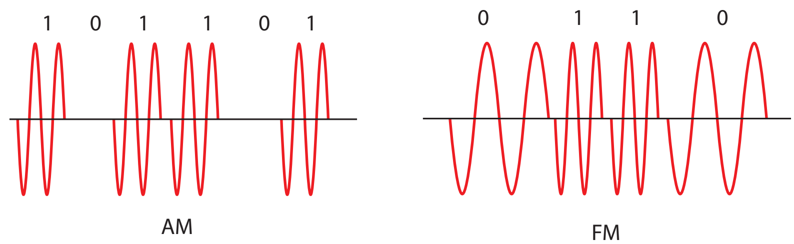


Figure 11.31 The difference between an amplitude modulated signal and a frequency modulated signal.

Amplitude modulation

The amplitude is modulated between two values as in Figure 11.31. This is different from AM radio where the amplitude varies continuously.

Frequency modulation

The frequency of the carrier wave is modulated between two frequencies, one representing 1 and the other 0. Computers can communicate with each other via wireless internet connections by using modulated radio waves.

Bandwidth

Remember that when a wave is modulated the wave changes from having one frequency to being made up of many frequencies, all added together to give the final wave shape. The digitally modulated waves have a wide bandwidth. This becomes a problem when using radio frequencies to transmit digital signals, since all the available bands are already used by radio stations.

Microwave

Microwaves are high frequency electromagnetic waves that when modulated enable a high rate of data transfer. Microwaves do not travel very far in the atmosphere and will not pass through concrete buildings. It is for this reason that the 'point to point' communication is necessary: the receiver must be in the 'line of sight' of the transmitter. The advantage is that no cables need to be laid. Microwaves are commonly used for wireless connections at home and in public places and in long range situations where there is a line of sight.

Note: The radio wave is always analogue, it is the information that is digital.

Light

Light has a frequency in the order of 10^{14} Hz. With such a high frequency a very high rate of data transfer is possible. Light can undergo amplitude modulation, in simplest terms, by switching on and off. The speed at which this can be done is currently the limiting factor to the maximum speed of data transfer using light. Other factors affecting the use of light are that the signal must be point to point, and it is affected by weather and other atmospherics. One way round this is to transmit the light signal in an optical fibre.

Satellite communication

Using high frequency modulated radio waves to transmit digital information would be an attractive possibility if it were not for the fact that the transmitter and receiver must be in direct line of contact with each other (line of sight). Therefore, to communicate over long distances, a chain of transmitters and receivers must be constructed. An alternative is to use a satellite to transfer the signal as in Figure 11.32.

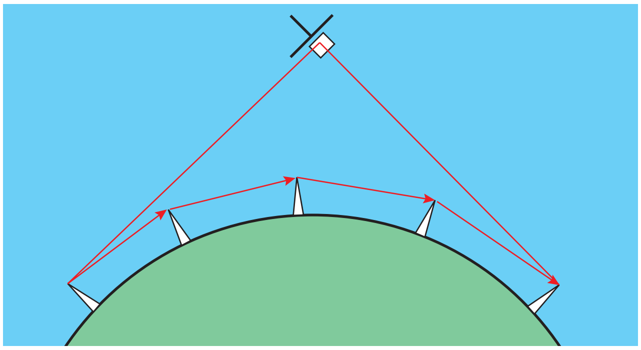


Figure 11.32 Signal transmitted from antenna to antenna and the alternative route via a satellite.

Geostationary satellites

A satellite is an object that travels around the centre of the Earth in a circular path. They are only useful for communication purposes if they stay in the same place. The Earth completes a full rotation once a day and, in order to stay above the same point on the Earth, a satellite must circle the Earth at the same rate and have the same axis of rotation. To make this possible the satellites are positioned above the equator. These are called *geostationary satellites*.

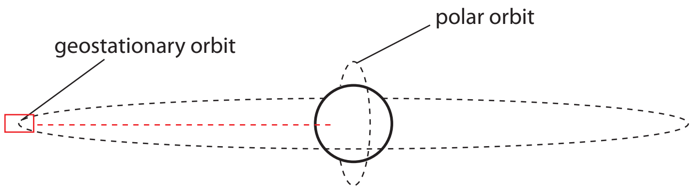


Figure 11.33 Comparison of polar and geostationary orbits. Polar satellites can orbit the Earth at a height of only 1000 km with a period of 2 hours. Geostationary satellites are always above the same point and orbit the Earth once a day at a height of 35 786 km.

Satellite TV

Geostationary satellites are used to transmit satellite TV stations. A TV broadcasting company transmits a microwave signal at a chosen uplink frequency (12–14 GHz) from a transmitter on Earth. The satellite receives the signal and sends it back to Earth at a different frequency, the downlink frequency. The reason for the change in frequency is to avoid interference with the uplink signal. On reaching the Earth, the signal is focused using a parabolic dish and decoded with a receiver. Since the satellite is far from the Earth, the signal will cover a large area. Satellite TV uses a compressed digital signal; the high frequency (12–14 GHz) gives a broad bandwidth, allowing a rapid transfer of data. The signal from one satellite can transmit up to 200 TV channels.

Kepler's Law

The square of the time for the satellite to complete one orbit (time period) is proportional to the cube of the orbit radius.

$T^2 \propto R^3$

This means that satellites in low orbits go around the Earth in less time than satellites in high orbits.

Polar satellites

A polar satellite orbits the North and South Poles. These satellites have a much lower orbit than geostationary ones and travel much faster. Using these satellites it is possible to take a signal from one place on the Earth and transfer it to another, this information could be stored and used later: every two hours you would receive the information for the next hour of TV programmes. Alternatively, if several

When a satellite is in polar orbit, it is possible to synchronize the orbit with the Sun, so that the satellite is always on the sunny side of the Earth. This is an advantage if the satellite is used for taking photos.

satellites are used, then the orbits can be arranged so that there is always one in view. This system is used to provide satellite TV for some areas of Russia that are too far north to be in line with geostationary satellites positioned over the equator.

Since these satellites are much closer to the Earth they are much cheaper to put in orbit and do not need to transmit such a powerful signal as the geostationary satellites.

Communication in remote areas

One of the benefits of satellite communication is that it is possible to reach parts of the world that are inaccessible with land-based systems. After a major earthquake or other natural disaster, the ground-based communication systems, such as telephone wires and radio antennae, are often destroyed, making communication impossible. Satellite systems are unaffected by such disasters and are invaluable to rescue teams working in the area.

International understanding

Land-based TV systems lie within national boundaries whereas satellite systems do not. Satellite systems, therefore, give many possibilities for cross cultural exchange. Viewers can watch news programmes with different perspectives, entertainment from other countries, or foreign language programmes that can aid language education. Communication leads to understanding and international communication leads to international understanding.

The Outer Space Treaty: Signed by 98 countries, this is a legal framework set up to control the use of outer space. To find more information, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 11.5.

Environmental concerns

Using satellites for communication reduces the need to lay cables and/or build many transmitter aerials on the ground. In this way it can be seen as an environmentally friendly means of communication. However, to put a satellite into orbit requires large amounts of fuel and so creates a large amount of pollution. In addition, launch pads have to be built in places away from large cities and may be in environmentally sensitive areas.

Colonizing space

There are ethical questions that need to be taken into consideration when using satellites for communication. Who owns space? Is space owned by the first country that put its flag there? Can anyone take over a piece of space for their satellite? What if two countries want to use the same piece of space, who decides who can use it?

Exercises

21 If the radius of a geostationary satellite is 42 000 km, what is the radius of a satellite with an orbit of 2 hours?

22 What is the speed of the satellite in Question 21?

23 If the radius of the Earth is 6400 km, what is the height of the satellite above the Earth?

Digitalizing everything

Most of the information we gain from our surroundings using our senses is analogue information; that is it continuously varies rather than happening in discrete steps. For example, when we use our sight to look at a scene, we see a continuous variation of colour and intensity, or when we listen to music we hear a continuous variation of frequency and loudness. We have seen in this chapter how it is possible to convert analogue to digital signals and how it is then possible to send this information with great speed along wires, optical fibres or via radio waves from one place to another. So it is possible to take pictures and sounds, digitalize them, process them with a computer then send them from one place to another at the speed of light. This is the basis of the internet.

The internet

From the early days of computing in the 1950s it was possible to send data from one computer to another, through a system called a *network*. The internet is far more extensive than this, for every computer can communicate with all the other computers that are connected. You may say this is just like the phone system. However, when you use the phone, the wires are connected by a system of switches called the telephone exchange. Once connected, your wires stay connected until you put the phone down; this is called a *circuit switching network*. The internet became a reality when *packet switching networks* became possible in the 1960s. Since the signal between computers is digital it can be split into packets, compressed and transmitted. When received it can be reassembled. As long as the bits are put back in the correct order, the signal can be mixed up with other data and even sent along different routes. So, unlike a telephone conversation where information is directed from one person to another, the information on the internet can be transmitted in bits. This packet switching system was developed by the military, so that vital information could be easily spread accross different locations rather than being stored on one vulnerable computer.

Tim Berners-Lee, inventor of the World Wide Web, who made his idea freely available, with no patent and no royalties due.

Advantages of the internet

Multiple communications

One of the main advantages of the internet is that you can communicate with many people at the same time. For example, sending an email to all the people on your address list, or posting an article for sale on eBay so anyone in the world can buy it.

Sharing information

With the internet came web pages which could be opened using a browser and contained information in the form of text and pictures. This has made web-based research, weblogs and social networking sites (such as Facebook and MySpace) accessible to everyone with an internet connection.

Business

It wasn't long after the initial development of the internet that businesses started to see the possible benefits. With its own internet site even a small business can advertise its products to the whole world. This is not only beneficial to the owners of the business but also to the consumers who can now find products, compare prices and even read consumer reports before deciding to buy something. The buying process can take place online, with Paypal for instance, and then a check on a bank balance can also take place online. With the advent of digitalizing and encoding signals, it has been possible for the internet to handle secure information such as credit card numbers and bank account details.

The internet was made possible by rapid advancement of technology but with its development have come many unforeseen problems. Are there some technologies that are best left undeveloped and who should decide whether to develop a new technology or not?

Global perspective
Since the internet had no central place or point of control, anyone could make a web page and share their experiences, knowledge, ideas and opinions. A certain knowledge of the programming language HTML was required but soon many individuals were mastering the procedure and making their own web pages. In recent years the process has been simplified so that any computer-literate individual can make their own web page or write a blog.

Some problems

One of the great things about the internet is the freedom it offers, for it isn't owned by anybody or run by anybody and anyone can use it. This of course, isn't quite true, since individuals must abide by the laws of the country they live in. However, the web is worldwide and different countries have different laws.

Needless to say, this freedom of expression has brought its own problems, some of which are outlined below.

Copyright infringement

As the speed of the signal increased it became possible to send not only pictures and text but also sound and video files. This made it possible to share the latest music and video; all it needed was for one person to buy a CD or DVD, copy it onto their computer and allow the world to copy it. Although this is illegal, it can be difficult to stop.

Pornography

In some countries around the world pornography is illegal. In the days before the internet, pornographic material would have to be smuggled from country to country. The internet opened up a whole new way of distributing pornography that was instantly exploited. Apart from the problem of monitoring the legality of content in different countries, there also came the problem of ensuring that the material was not viewed by young children. The 'Do not enter if under 18' signs probably did little to solve this problem.

Extreme views

An individual with extreme views living in a small community may not find many people who share those views. However, with the world as an audience through the internet it is very likely that groups of like-minded people will be able to come together.

Gambling

The ability to exchange money online opened up the possibility of online gambling. Here there is often no regulation, no doormen who check identity and refuse entry to those who are under age or have had too much to drink. Gambling online at home can take place 24/7 with no restrictions.

Plagiarism

Students writing essays (and authors writing books) have instant access to millions of sources of information; they also have access to completed essays. Plagiarism is nothing new but has become a bigger problem through the internet.

Spam

The ability to send emails to millions of people at the same time is open to abuse by those pushing material and information on to people who don't want it.

11.5 Electronics

Assessment statements

F.5.1 State the properties of an ideal operational amplifier (op amp).

F.5.2 Draw circuit diagrams for both inverting and non-inverting amplifiers (with a single input) incorporating operational amplifiers.

F.5.3 Derive an expression for the gain of an inverting amplifier and for a non-inverting amplifier.

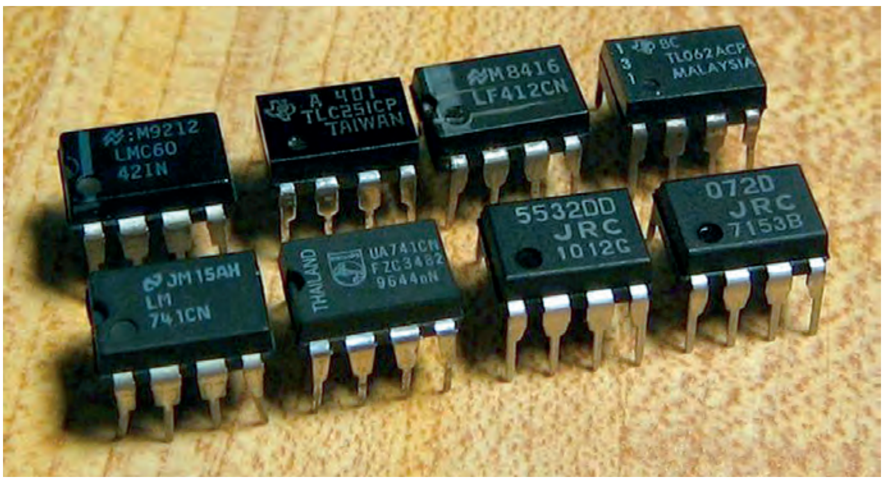
F.5.4 Describe the use of an operational amplifier circuit as a comparator.

F.5.5 Describe the use of a Schmitt trigger for the reshaping of digital pulses.

F.5.6 Solve problems involving circuits incorporating operational amplifiers.

The operational amplifier

An amplifier is a device that takes a signal and makes it bigger. For example, an MP3 player gives out a very small p.d. that produces sound you can hear using headphones but not in your whole room. If you connect the signal from the MP3 player to an amplifier, then the signal can be made big enough to drive a large pair of speakers. The amplifier needs energy from a supply, such as the mains, to power the speakers. However, an operational amplifier, or op amp, can be made to do more than amplify. It can be wired in different ways to carry out functions that are the basis of many devices used in digital electronics.



Op amp chips.

The actual op amp is a small black rectangle called a *chip*. It has 8 numbered connecting legs although we will only use 5 of them. Inside the chip is a complicated circuit with many transistors and resistors etched onto a thin piece of silicon. Fortunately, you do not have to understand all the internal parts; we are only concerned with what happens to the output when we change the input.

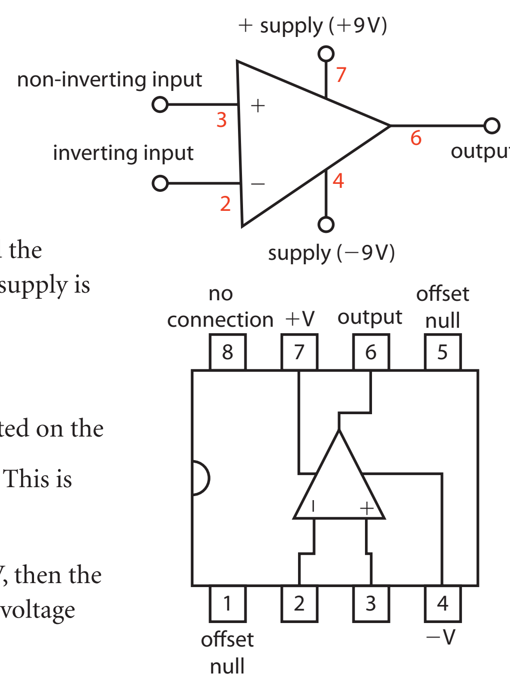


Figure 11.34 The op amp symbol. The numbers show how the 5 terminals of the op amp are connected inside to the legs on the chip.

In Figure 11.34, the inputs show where the signal is fed into the amp and the output shows where the amplified signal comes out. In this diagram the supply is shown to be 9 V. This is the maximum value that the output can be.

Open loop gain

If a very small p.d. is applied between the inputs, a large potential is created on the output. The ratio $\frac{V_0}{(V_+ - V_-)}$ is called the *open loop gain* of the amplifier. This is typically about 1 000 000 .

This would mean that if the p.d. between the input terminals were 0.01 V, then the output potential would be 10 000 V. This is not possible since the supply voltage (9 V) can not be exceeded.

In an ideal op amp, the open loop gain is infinite.

Inverting and non-inverting

If V_+ is bigger than V_- the output is +9 V. Since both are +ve, the +ve terminal is called non-inverting.

If V_- is bigger than V_+ the output is −9 V. Since these have opposite signs, the −ve terminal is called the inverting input.

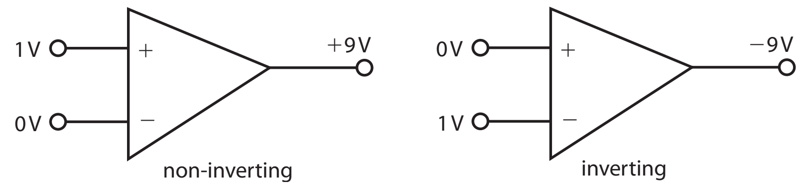


Figure 11.35 Inverting and non-inverting amplifiers (note: supply not shown).

Input impedance

This is effectively the resistance that would be measured if a meter were placed across the input terminals. For an ideal op amp this is infinite, therefore no current flows into the inputs.

Output impedance

The output of the op amp is like a 9 V battery: it can be used to power loudspeakers, light bulbs and other devices. The output impedance is equivalent to the internal resistance of the battery. In an ideal op amp this is zero. This means that it can deliver a lot of current. To get a current to flow through a load resistor there must be a p.d. across it; this is achieved by connecting the resistor between the output and a wire at 0 V. This is called an *earth*.

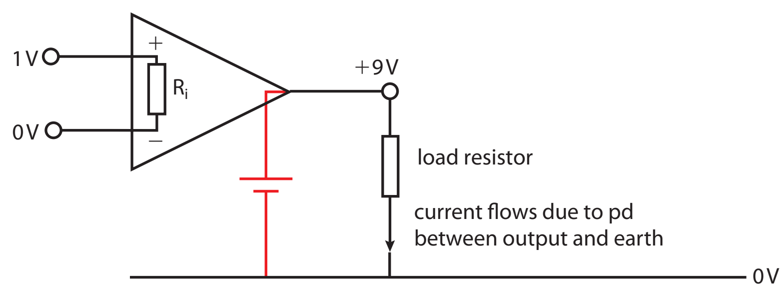


Figure 11.36 The input impedance is represented by R_i. The output is like a battery with zero internal resistance.

The ideal op amp

The ideal op amp can be summarized as having:

- infinite input impedance
- infinite gain
- zero output impedance.

No op amp is really like this but this assumption makes it easier to understand how the following applications work.

Exercises

24 An op amp with an open loop gain of 1 000 000 is connected to a 9 V source. What will the output be if:

(a) the potential of the non-inverting input is 5 μV?
(b) the potential of the non-inverting input is 5 V?
(c) the potential of the inverting input is 5 V?

Amplifiers

Using an op amp on its own is not very useful for amplifying music. If the open loop gain is 1 000 000, then once the signal goes above 0.000 009 V the output will be a constant 9 V. An amplifier must be able to produce a signal that is the same shape as the input but with larger amplitude.

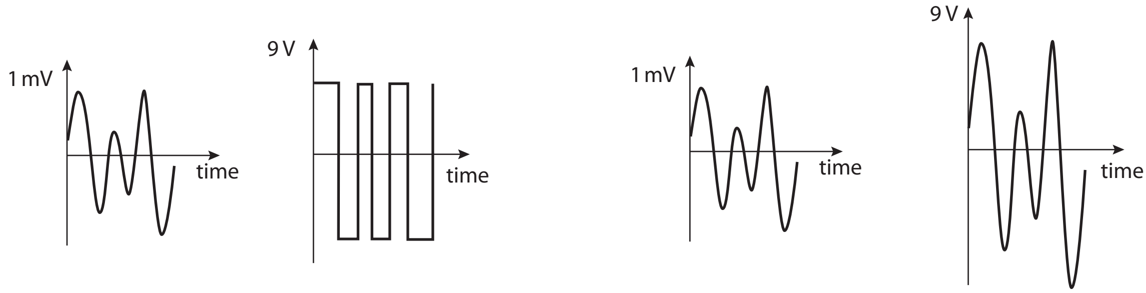


Figure 11.37 Comparison of a signal amplified open loop and a signal properly amplified.

The non-inverting amplifier

Consider the circuit for a non-inverting amp shown in Figure 11.38. We know that if the difference between V_+ and V_- is more than 9 μV the output will be a constant 9 V. Therefore, we can say that the two inputs are about the same:

$$V_+ = V_- = V_{\text{input}}$$

This means that the potential in the middle of R_1 and R_2 is also V_{input}

If the current through this combination is I, then applying Ohm's law to the combination R_1 and R_2 gives $I = \dfrac{V_{\text{output}}}{R_1 + R_2}$

Applying Ohm's law to R_2 alone $I = \dfrac{V_{\text{input}}}{R_2}$

Equating gives $\dfrac{V_{\text{output}}}{R_1+R_2} = \dfrac{V_{\text{input}}}{R_2}$

So $\dfrac{V_{\text{output}}}{V_{\text{input}}} = \dfrac{R_1 + R_2}{R_2} = 1 + \dfrac{R_1}{R_2}$

This is the gain of the amplifier.

Simplifications

You may wonder how there can be an output if the two inputs are equal.

Well of course there wouldn't be, this is a simplification to enable us to be able to calculate the gain.

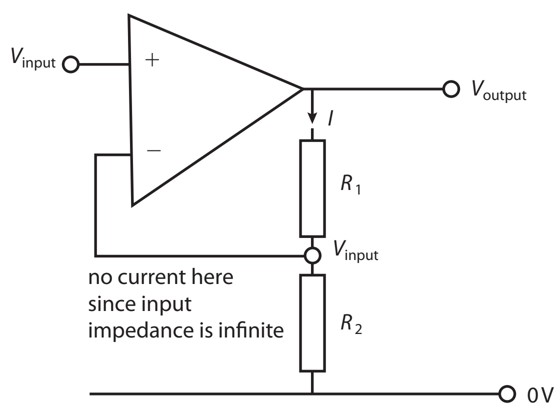


Figure 11.38 A non-inverting op amp.

Exercises

25 Referring to the non-inverting amplifier in Figure 11.38 (the supply is 9 V):
If $R_1 = 10\ \text{k}\Omega$ and $R_2 = 1\ \text{k}\Omega$:
(a) What is the gain of the amplifier?
(b) If the input is 0.5 V what will the output be?
(c) Calculate the current through R_1.
(d) What is the p.d. across R_1?

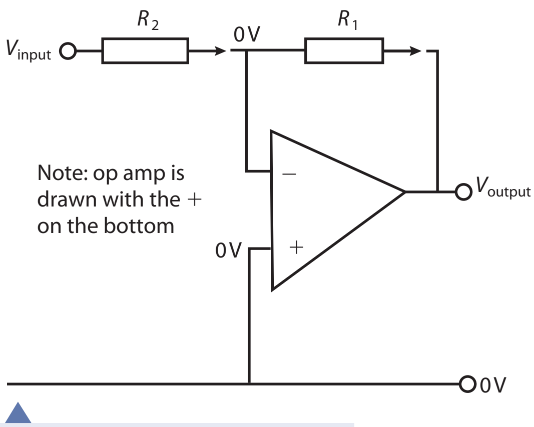


Figure 11.39 The inverting amplifier.

The inverting amplifier

Now consider an amplifier that gives an inverted signal as shown in Figure 11.39. Again the difference between V_+ and V_- must be very small, so let's say they are equal. Since V_+ is connected to earth then V_+ and V_- are equal to 0 V.

$V_+ = V_- = 0$

The potential at a point between R_1 and R_2 is therefore also 0 V.

Applying Ohm's law to R_1 gives $I = \dfrac{V_{input}}{R_2}$

Applying Ohm's law to R_2 gives $I = \dfrac{V_{output}}{R_1}$

Equating gives $I = \dfrac{V_{input}}{R_2} = \dfrac{V_{output}}{R_1}$

$$\frac{V_{output}}{V_{input}} = \frac{R_1}{R_2}$$

Exercises

Referring to the inverting amplifier in Figure 11.39 (the supply is 10 V):

26 If $R_1 = 5\,\text{k}\Omega$ and $R_2 = 1\,\text{k}\Omega$

(a) Calculate the gain of the amplifier.

(b) If the input is 1 V what will the output be?

(c) Calculate the current through R_2.

(d) Calculate the p.d. across R_2.

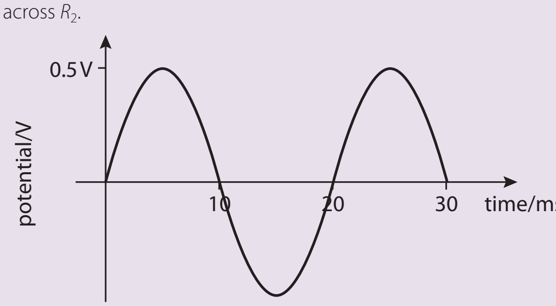


Figure 11.40

27 If the signal shown in Figure 11.40 is applied to the input of the amplifier in Question 26

(a) What will the amplitude of the output be?

(b) What will the frequency of the output be?

(Remember frequency $= \dfrac{1}{\text{time for a complete cycle}}$.)

28 Draw the output obtained if the rising potential shown in Figure 11.41 is applied to the input of the amplifier in Question 26 .

29 If an inverting amplifier has a gain of 10, what will the value of R_2 be if $R_1 = 5\,\text{k}\Omega$?

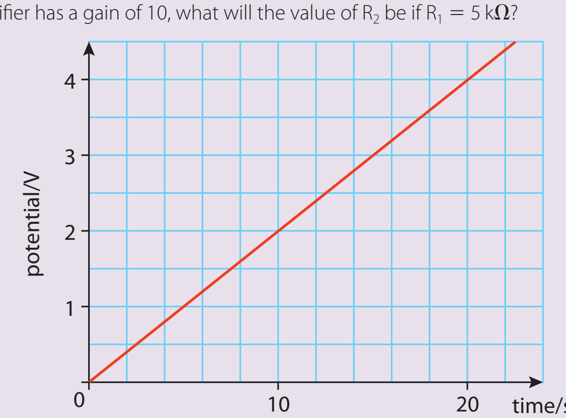


Figure 11.41

Comparator

A comparator compares two potentials. If the potential on the + (positive) input is higher, then the output is +9 V and if the − (negative) is higher, the output is −9 V.

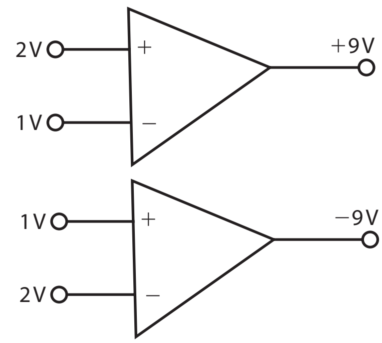


Figure 11.42 An op amp as a comparator

A simple fire alarm

Figure 11.43 shows a simple fire alarm that uses a thermistor to detect when the temperature goes above a certain value.

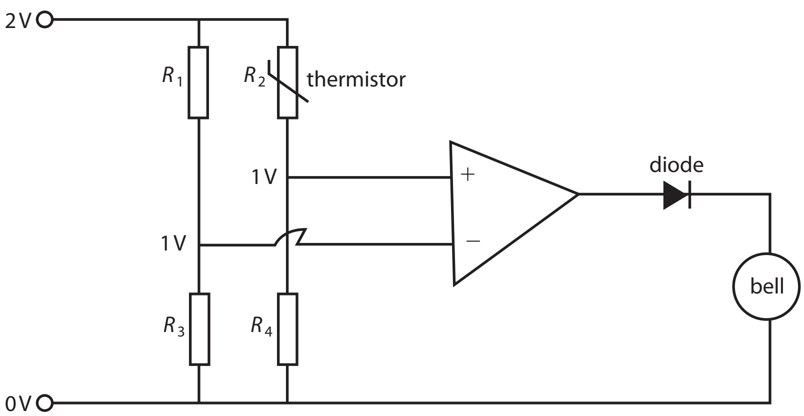


Figure 11.43 The circuit for a simple fire alarm includes a thermistor. This is a semi-conducting device with a resistance that goes down when its temperature increases. It also includes a diode which only allows current to flow in the direction of the arrow.

At normal room temperature all the resistances are equal ($R_1 = R_2 = R_3 = R_4$) so both inputs equal 1 V and the output is zero. If the temperature of the thermistor increases then its resistance gets less, resulting in a smaller p.d. across it. The potential to the + input will therefore rise, resulting in an output of 9 V. This will cause the bell to ring. The diode prevents the bell ringing if the thermistor gets cold.

This circuit is very sensitive, so the bell will ring even if the temperature rises only a small amount. The input can be set higher by varying the potential divider and so making make the bell ring only when there is a fire.

Exercises

30 An op amp can be used to ring a bell if the temperature of a room falls below a certain value. The circuit is similar to the circuit in Figure 11.43 and uses a thermistor whose resistance varies with temp according to Table 7.

Temp/°C	Resistance/Ω
17	146.6
18	139.6
19	133.0
20	126.7
21	120.8
22	115.2
23	109.8

Table 7

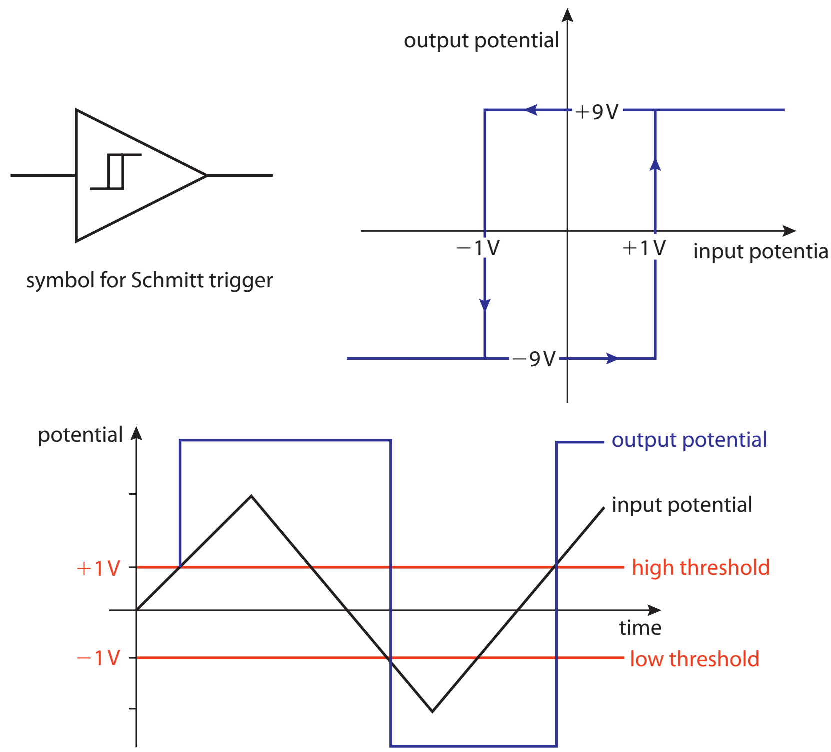
(a) Suggest a way the circuit could be changed (involving the diode) so the circuit performs this task.
(b) What are the values of R_1, R_3 and R_4 so that the bell rings if the temperature drops below 18° C?
(c) If the temperature falls to 17° C, how much current will flow through the thermistor?
(d) At 17° C what will the potential at the + input be?
(e) At 17° C what will the output of the op amp be if the supply voltage is 9 V and the gain is 1 000 000?

Remember
R_2 and R_4 are a potential divider so the potential divider equation
$V_{out} = V_{in}\dfrac{R_4}{R_2 + R_4}$
can be used.

Schmitt trigger

A Schmitt trigger is a type of comparator that switches to a high output when the input is above a given value (upper threshold) and only switches to a low output when the input gets lower than another low value (lower threshold). For example, if the thresholds are +1 V and −1 V, then if the input is above 1 V the output is +9 V until the input drops below −1 V when the output changes to −9 V. When the input is between +1 V and −1 V, the output is +9 V if it is on the way down and −9 V if on the way up. In this way the comparator remembers what its previous state was; this is called *hysteresis*.

Figure 11.44 The input and output of a Schmitt trigger and the relationship between these shown on the hysteresis curve.



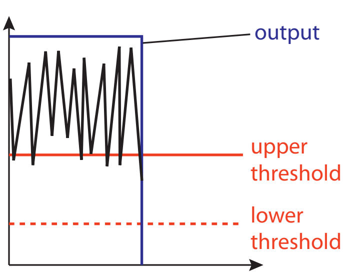


Figure 11.45 The noise on this digital signal is represented by the rapid variation on the signal. The Schmitt trigger switches the output to high when the input goes over the upper threshold and stays high until it goes below the lower threshold and so removes the noise.

Uses of the Schmitt trigger

One of the applications of a Schmitt trigger is for reshaping digital signals, especially to remove noise (see Figure 11.45).

Exercises

31 The potential in Figure 11.46 is connected to the input of a Schmidt trigger with an upper threshold of +1 V and a lower threshold of −1 V.

Draw a graph of the output if the power supply to the op amp is 9 V.

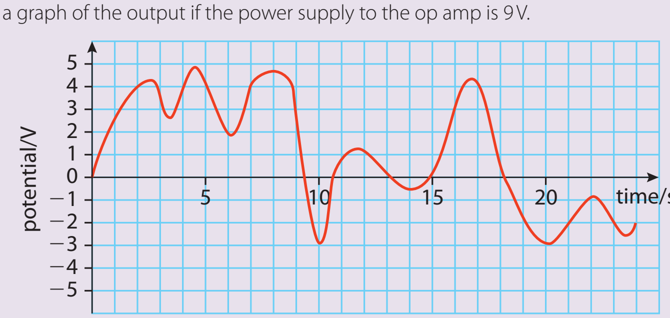


Figure 11.46

11.6 The mobile phone system

Assessment statements

F.6.1 State that any area is divided into a number of cells (each with its own base station) to which is allocated a range of frequencies.

F.6.2 Describe the role of the cellular exchange and the public switched telephone network (PSTN) in communications using mobile phones.

F.6.3 Discuss the use of mobile phones in multimedia communication.

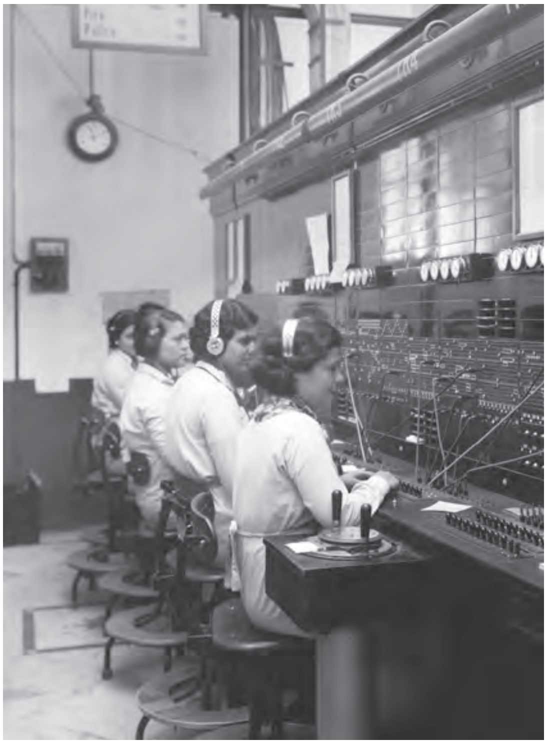
F.6.4 Discuss the moral, ethical, economic, environmental and international issues arising from the use of mobile phones.

The mobile phone system

The mobile phone is a combination of two technologies: radio and telephone. These have been in use since the 1880s but the progress in the technologies is in the size and functionality of modern mobiles. For instance, a mobile can be a camera, an MP3 player, a calendar and an alarm clock as well as a phone. What are the developments that have made this possible?

The land phone system

The land-based telephone system today is very complicated and can perform many functions. However, the principle (and a lot of the wiring) is still the same as in the early days. In the early 20th century the phones in each town were connected to an operator and the operators in each town were connected to each other. When the phone was picked up a bulb would light up in a plug on the operators' switchboard and an operator would plug their headset into your line and ask who you wanted to speak to. If it was to someone in the same town then the operator would simply connect the two phone lines together but if it was further away then the operator would contact a second operator who would connect you. A long-distance call could involve many operators. Today the system has been replaced by electronic switches and when a number is dialled each digit is recognized and the caller is connected to the receiver. This can involve local, national and international exchanges. The complete system is called the *public switched telephone network* or PSTN.

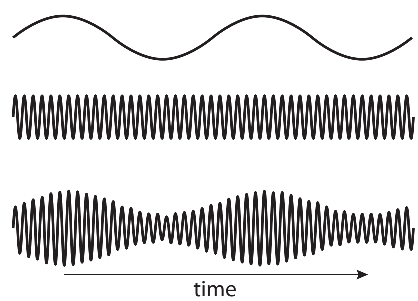


Telephone operators routing calls at a local exchange in the early 20th century.

Radio telephone communication

Radio uses low frequency electromagnetic waves to transmit signals, rather than electric current in wires as used by telephone. Radio waves spread out in all directions, so caller and receiver cannot be connected together using switches and wires. Instead, each caller and receiver choose a certain frequency and the sound is then superimposed on this carrier frequency and transmitted. Some of the drawbacks with this system are that anyone may tune in to a conversation, and the transmission has a fairly short range unless you use a very big antenna and a powerful transmitter. In fact, it has been possible to connect a radio to the phone system in this way for many years. Each connected city had an antenna that could transmit and receive the signals from each radio phone and each phone had a selection of some 25 different frequency channels. So only 25 people could use the radio telephone system at any one time, and they had to be near a radio antenna! Also, the phones were huge as they had to have a powerful transmitter that needed a large battery.

Figure 11.47 Sound can be superimposed on a carrier frequency by changing the amplitude of the carrier to match that of the sound. This is called amplitude modulation.



The cell phone

Modern mobile phones are radio transmitters in the same way as the old radio phone systems, but operate using *cells*, hence the name *cell phone*. As we have seen, the main problems with a radio telephone communication system are the need for a powerful transmitter and the limit on the number of users at any one time. If each city had 100 small antennae then this would increase the number of users and also take away the need for mobiles to have powerful transmitters. It would mean that if you were moving around across several cells whilst talking on the phone (the whole point of a mobile phone) you would need to be switched from one transmitter to another. To prevent a break in the conversation, this switching would need to be done quickly and automatically. It is only recently that computer technology has been able to perfect the switching process, enabling cell phones to become very much part of modern life.

A cell is the local area covered by one of the short-range transmitters in a cellular telephone system.

The original mobile phones received an analogue signal via a radio wave but the switching of the signal was digital. The modern phones use a higher frequency radio signal which has a digital signal superimposed on it. In this way many more channels can be opened enabling several hundred communications to take place simultaneously through the same antenna.

How the cells are arranged

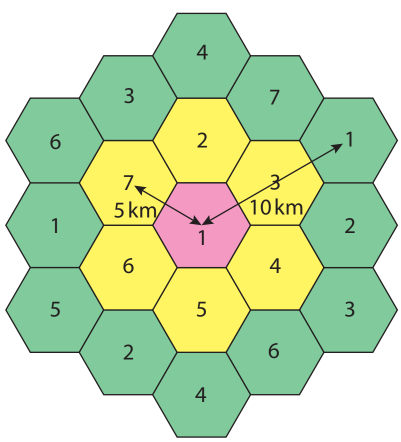
A city is split up into many small units, or cells, of hexagons each with an area of about 26 km^2 to enable a mobile phone to get coverage of the whole city. This is because a mobile has quite a weak transmitter and only has an operating distance of about 7 km (see Figure 11.48).

A mobile phone company is only allowed to use a limited range of frequencies (between 824 MHz and 894 MHz) so that phone calls don't interfere with radio transmission. This range will give the company about 800 channels that it can use.

Each conversation must use two channels (full duplex); otherwise it would not be possible for the two people to talk at the same time. This limits the channels to 400. The seven cells that are closest to each other (the pink and yellow) cannot use the same bands, so the 400 available are split into 56 channels per cell. In this way the channels are organized so that no matter where you are, you cannot receive the same signal from two transmitters/antennae. The cells using the same channels (see Figure 11.48) are arranged so they are not next to each other. One advantage of this system is the ease with which it can be expanded to cater for more users. More channels can be added by simply reducing the size of the cells.

Receiving a call

When you switch on your mobile it sends out a signal, this isn't one of the 400 conversation bands but another band that has been reserved for this purpose. All the antennae in range in the cells will receive this signal but only the one that gets the strongest signal responds, and then only if you are a member of the right network. If you move to another cell then another antenna gets the strongest signal and knows to direct any calls to this new position.



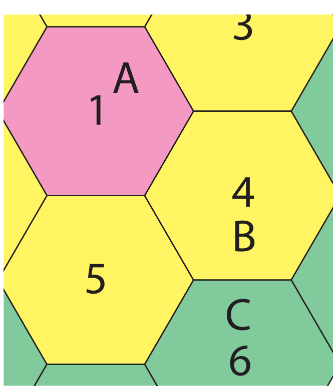


Figure 11.48 In this arrangement of cells a phone in a pink cell is close enough to receive signals from all the transmitters (antennae) in the yellow cells but not the green ones. The pink and yellow cells cannot, therefore, use the same channels. The cells with the same number use the same channels and are arranged so they are not next to each other. If you start in position A you will receive a signal from pink cell 1 that will inform you that you are connected to the network. Any calls will be directed to pink 1 but if you now walk from A to B the signal to yellow 4 will become the strongest and the call will now be transferred to yellow 4.

Tracking

For the digital switching system to operate, the transmission towers/antenna of the cell you are in and all the cells that you might go into need to be receiving your signal. The antenna with the strongest signal is the one which relays the phone call to you, but the others 'know you are there'. From the strength of the signal it is possible to calculate how far you are from each of the antennae and then some simple geometry will define your position.

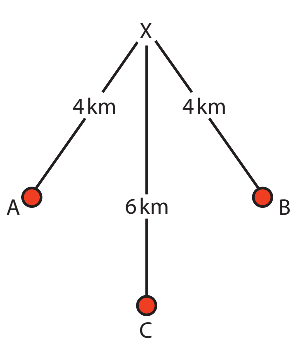


Figure 11.49 If you are 4 km from A, 4 km from B and 6 km from C you must be at point X.

3G

The latest cell phones are known as third generation or 3G. These use not only digital switching but also a digital signal to enable the transfer of more than just conversations. 3G phones can open web pages, search the internet, download video files and pictures. Since the phone itself is a digital device, it can also process data and communicate with a computer – in fact it is closer to a computer than a telephone.

Advantages and disadvantages of the cell phone

Mobile communication

The obvious advantage of the cell phone is the possibility of communicating on the move. Before the introduction of the mobile phone you would have to be at home or in an office if you wanted to communicate with someone, unless you were one of the few people able to afford a radio phone (for conversations only, not text). When travelling, a phone box was the most common way to communicate (providing you had some coins in your pocket).

Not only can you talk whilst on the move, but digital technology has opened the doors for sending texts as a cheap and efficient way of communicating, or pictures, photos and video clips.

Tracking

The use of mobile phone tracking to catch criminals is often used in crime series on the television. This gives the impression that the police are doing it all the time. Maybe reality isn't quite the same.

It has been possible to track a mobile phone user's movements since the birth of cellular technology, but this was not the original purpose of the technology and so this facility has only been available to the police and emergency services until recently. Now it is available to the public and for a small fee you can receive information on the position of a mobile. It is easy to see how the police may have tracked criminals and how emergency services may have rescued lives. However it is not difficult to see how this can be abused.

Bringing people together

It is a relatively simple and cheap process to set up a cell phone network compared to running wires from house to house. This has meant that it has been possible to introduce mobile phones into areas that have not had communication before.

Mobile phone etiquette

Following close on the heels of the development of mobile phone technology has been the development of the social etiquette for its use. It has become accepted that there are places where the use of a mobile is not desirable and many stores, theatres and other public places now have signs asking the public to turn off their phones before entering. This avoids irritating ring tones, loud conversations, disturbances in hospitals, and interruptions to concert and theatre performances.

Social alienation

The mobile phone is a method of communication so it might seem strange that its use could lead to social alienation. However, there are certain things that are easier to say within a text message than face to face. Does the use of mobile phones and texting reduce the amount of time that people spend talking to each other?

It is interesting to consider whether groups of young people sitting together and texting others rather than talking to each other has resulted in an increase in communication by including people who can't be there, or resulted in an decrease in communication between members of the group? Whatever the answer, the mobile phone has certainly changed the way we communicate.

It is very difficult to collect scientific data on the health effects of mobile phones. As it is not ethical to expose humans to higher and higher levels of high frequency radio waves until they develop symptoms, the data on health and mobiles comes from people who have become sick and have used mobile phones. It is difficult to say if the mobile phones have been the cause of the illness.

Health

There has been a lot of talk in the media about the possible dangers of using a mobile phone. Reports mention possible links between exposure to electromagnetic radiation and cancer. However, concentrating on the physics, when the high frequency radio waves are absorbed by the human body, they will cause a slight rise in temperature, but the energy is not high enough to excite the atomic electrons that would be necessary to cause chemical changes.

The greatest health risks for mobile phone users are those related to the increased stress of being contactable 24 hours a day.

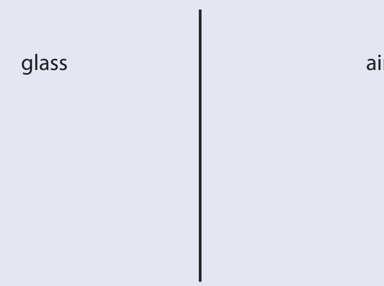
Environmental problems

To set up a cell phone network requires the building of antennae. This is not normally a problem in cities, where they can be put on top of existing buildings or even hidden inside church spires. However, in rural areas, the antennae are unsightly and often unwanted. Another environmental problem associated with this rapidly changing technology is the disposal of phones. Before the birth of mobile technology, when you bought a house, the phone came with it, and the life of the phone could be thirty or forty years. There were of course new styles of phone on the market but since few people would ever see your phone, it wasn't so popular to keep up with the new designs. The mobile is not just a phone but has become a fashion accessory and a status symbol. New models come out every week and old models look dated after a year. The life of a mobile is on average about 18 months, which means in a country the size of the UK around 15 million phones are discarded each year. Mobiles phones contain cadmium, rhodium, palladium, beryllium and lead solder, which are all highly toxic. For this reason, phones should be recycled and not thrown into the bin.

A mobile phone antenna shaped like a tree to make it blend into the environment!!

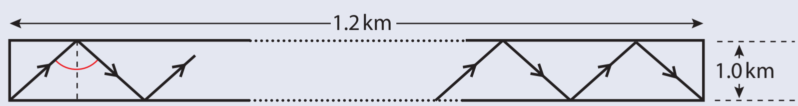
Practice questions

1 This question is about refractive index and critical angle.
The diagram below shows the boundary between glass and air.



(a) On the diagram, draw a ray of light to illustrate what is meant by critical angle. Mark the critical angle with the letter "c". (3)

A straight optic fibre has length 1.2 km and diameter 1.0 mm. Light is reflected along the fibre as shown below.

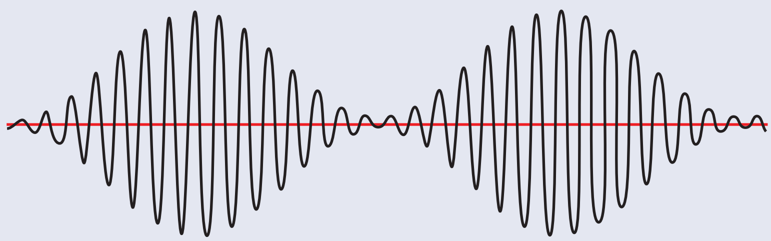


At each reflection, the angle of incidence is equal in value to the critical angle. The refractive index of the glass of the fibre is 1.5.

(b) Deduce that the length of the light path along the optic fibre is about 1.8 km. (4)
The speed of light in the fibre is $2.0 \times 10^8\ \text{m s}^{-1}$.

(c) Calculate the time for a pulse of light to travel the length of the fibre when its path is
(i) along the axis of the fibre (1)
(ii) as calculated in **(b)**. (1)

2 This question is about amplitude modulated waves.
The sketch below shows the voltage–time graph for an amplitude modulated wave.



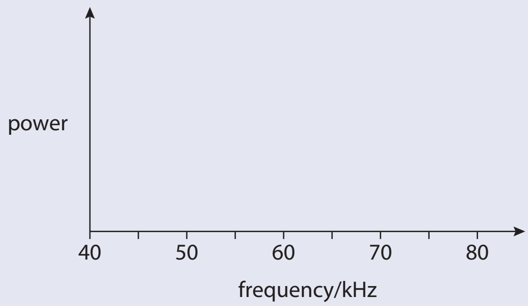
The frequency of the carrier signal is 60 kHz.

(a) Calculate the frequency of the information signal. (2)

(b) Estimate the modulation index. (1)

(c) If the maximum amplitude of the modulated signal is 2V, what is the amplitude of the information signal? (1)

(d) On the axes below, draw the power spectrum of the signal.



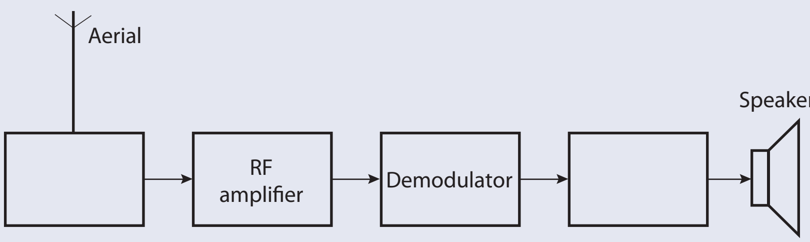
(2)

3 An AM radio station is transmitted with a carrier frequency of 198 kHz. The signal has a range from 100 Hz to 3.5 kHz. (2)

(a) Calculate the bandwidth of the signal. (2)

(b) Sketch the frequency spectrum of the signal. (2)

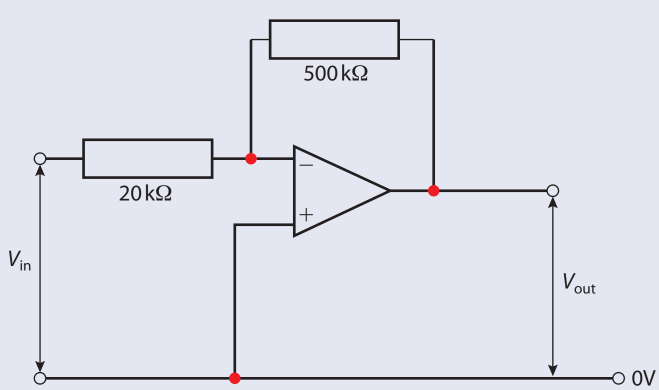
(c) Complete the system diagram of a simple AM radio receiver.



(d) Explain the function of the RF amplifier. (1)

(e) Explain the function of the demodulator. (1)

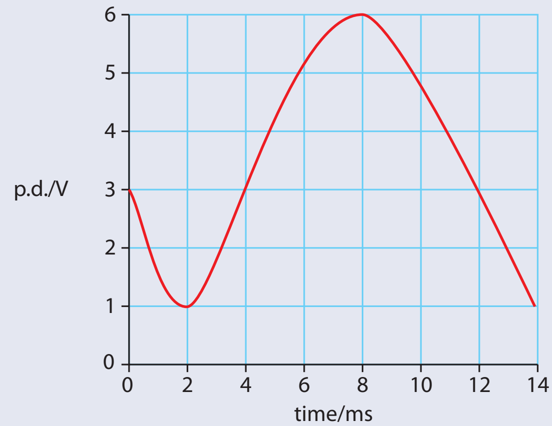
4 Consider the following OP amp circuit.



(a) Is this an inverting or non-inverting amplifier? (1)

(b) Calculate the gain of the amplifier. (2)

5 **(a)** The graph below shows an analogue signal.

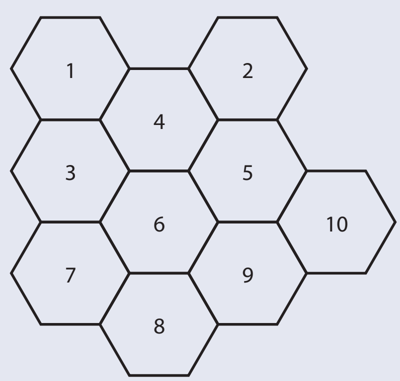


This signal is sampled at 500 Hz. Fill in the table below giving the voltage level and equivalent 3-bit binary code each 2 ms. (4)

Time/ms	Voltage level	Binary
0		
2		
4		
6		
8		
10		
12		

(b) Give one advantage of storing music in digital rather than analogue form. (1)

6 This question is about the cell phone network shown in the diagram below.



(a) If you are in cell 6 which transmitters can you receive signals from?
The frequency of the carrier signal is 60 kHz. (1)

(b) Explain why adjacent cells cannot use the same frequency. (1)

(c) Describe what happens if you walk from cell 6 to cell 5 whilst you are talking on the phone. (1)

(d) Give one environmental problem related to cell phone use. (1)

12 Electromagnetic waves (Option G)

12.1 The nature of electromagnetic (EM) waves and light sources

Assessment statements

G.1.1 Outline the nature of EM radiation.
G.1.2 Describe the different regions of the EM spectrum.
G.1.3 Describe what is meant by the dispersion of EM waves.
G.1.4 Describe the dispersion of EM waves in terms of the dependence of refractive index on wavelength.
G.1.5 Distinguish between transmission, absorption and scattering of radiation.
G.1.6 Discuss examples of the transmission, absorption and scattering of EM radiation.
G.1.7 Explain the terms *monochromatic* and *coherent*.
G.1.8 Identify laser light as a source of coherent light.
G.1.9 Outline the mechanism for the production of laser light.
G.1.10 Outline an application of the use of a laser.

The origin of EM radiation

If your physics teacher were to bring a charged object into class, it would make your hair stand on end, not because it is frightening but because your hairs would be repelling each other. The way we explain how this can happen, even though the object is not touching you, is by defining an electric field. This is a region of space where we feel the effect of electric charge. Students at the back of the room wouldn't feel such a large effect since the field gets less with distance. If the charge is now moved away from the students everyone will feel a change, the field will be weaker and everyone's hair will go down. This spreading out of disturbance is like the spreading out of a water wave when a stone is dropped into a pool of water, except the disturbance is of an electric field not the surface of water.

When a charge is moved, a second field is created, a magnetic field. (This is what happens when a current flows through a wire.) The direction of the magnetic field is perpendicular to the electric field. The result is that when a charge is moved, a changing electric and magnetic field spreads out. If this disturbance meets a different medium we find that it is reflected and refracted. When it passes through a small opening, it is diffracted and two different disturbances interfere. These are the properties of a wave so this is called an *electromagnetic wave*.

Scottish physicist James Clerk Maxwell (1831–1879) who discovered the magnetic field associated with a changing electric field.

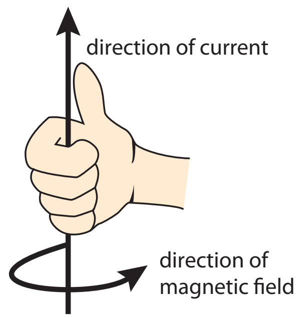


Figure 12.1 The right-hand grip rule. If a wire is gripped as shown, the fingers show the direction of the magnetic field and the thumb shows the current direction.

Creating an electromagnetic wave

An electromagnetic wave can be created by passing an alternating current through a wire as shown in Figure 12.2. Waves created in this way are called *radio waves*. James Maxwell found that it was not the moving charge that caused the magnetic field but the changing electric field that was causing the charge to move. This explains how electromagnetic waves can travel through a vacuum: the changing fields induce each other. Maxwell also calculated that the speed of the wave in a vacuum was approximately $3 \times 10^8\ \mathrm{m\,s^{-1}}$. This value was about the same as the measured value for the speed of light, so close in fact, that Maxwell concluded that light was an electromagnetic wave.

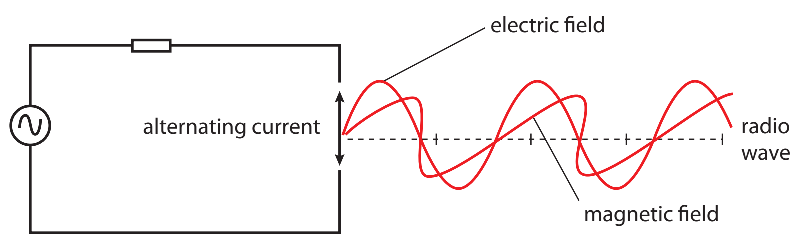
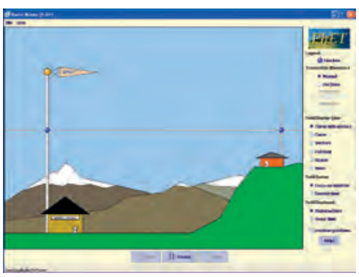


Figure 12.2 An alternating current produces a radio wave.



The PhET simulation 'Radio Waves' shows the components of a wave. To view, visit www.heineman.co.uk/hotlinks, enter the express code 4426P and click on Weblink 12.1.

How light is produced

The difference between light and radio waves is the frequency. Light waves have a much higher frequency than radio waves. However, it is not possible to produce light by simply moving a charge up and down very fast, as it is not possible to change the direction of charge quickly enough.

Light comes from the individual atom. Atoms contain electrons that can exist in different energy levels. When an electron changes from a high energy level to a low one it gives out energy in the form of electromagnetic radiation. The frequency of the light, f, is related to the change in energy ΔE by the equation:

$$\Delta E = hf$$

where h = Planck's constant $6.6 \times 10^{-34}\ \mathrm{m^2\,kg\,s^{-1}}$

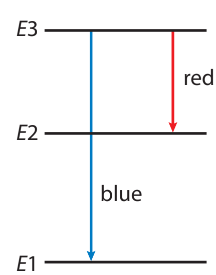


Figure 12.3 Since there are many electron energy levels in an atom this leads to the emission of light with many different frequencies, each frequency corresponding to a different colour.

Even higher frequencies

Electron energy levels are in the order of 10 eV. This is $10 \times 1.6 \times 10^{-19}$ J.

Using the formula $\Delta E = hf$, an energy change of 10 eV will give rise to light with a frequency of 2.42×10^{15} Hz. However, EM radiation with much higher frequency, over 10^{20} Hz, does exist. This would need an energy change in the order of MeV, much greater than electron energies. Radiation with such high energy comes from the nucleus.

Exercises

1 Calculate the frequency of light emitted when an electron changes from an energy of 10 eV to 6 eV.

2 An atom has electrons that can exist in 4 different energy levels, 10 eV, 9 eV, 7 eV and 2 eV. Calculate:
(a) the highest frequency radiation that can be produced
(b) the lowest frequency radiation.

3 What energy change would be required to produce EM radiation with a frequency of 1×10^{18} Hz?

The electromagnetic spectrum (EM spectrum)

Wavelength

The frequency of electromagnetic waves can range from almost 0 Hz up to 10^{20} Hz. The speed of all electromagnetic radiation is $3 \times 10^8\,\mathrm{m\,s^{-1}}$. Using the formula $v = f\lambda$ we can calculate the wavelength of the waves.

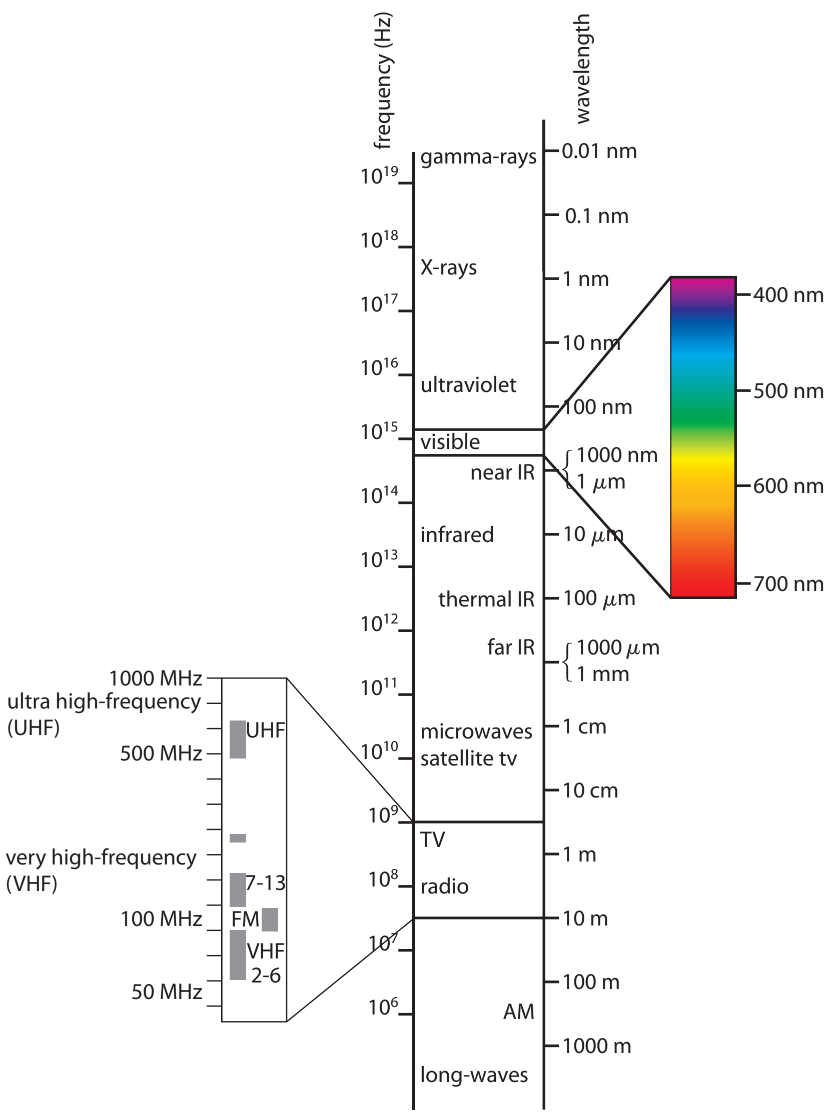


Figure 12.4 The electromagnetic spectrum. Waves can be classified in terms of their wavelength. Each range of wavelength has a different name, different mode of production and different uses.

● **Examiner's hint:** Note that the regions are not clearly separated. For example, there is considerable overlap between X-rays and gamma rays.

Why do humans use the visible range of frequencies to see?
With all these different types of EM radiation you may wonder why we use the frequencies we do. Well, most of the light we use to see comes from the Sun. This contains many more frequencies than just the visible light; however most of the frequencies are absorbed when the radiation passes through the atmosphere. It's also true that if we used radio to see then we would have to have antennae instead of eyes and that wouldn't look very attractive!

Worked example

What is the wavelength of green light with a frequency of 6×10^{14} Hz?

Solution

Rearranging $v = f\lambda$ gives $\lambda = \frac{v}{f}$

so the wavelength of green light $= \frac{3 \times 10^8}{6 \times 10^{14}}$ m

$\boldsymbol{\lambda = 500 \times 10^{-9}}$ **m**

Radio waves

Radio waves are produced from an alternating current in a tuned electrical circuit. Radio waves are used in communication; they are split into smaller subdivisions according to frequency. High frequency waves can carry more information per unit time than low frequency waves, so they are used for the rapid transfer of information required by satellite TV and the internet. Low frequency radio is used by the traditional radio stations.

Microwaves

Microwaves are produced by oscillations of electrons in a vacuum. The EM wave produced resonates in a hollow metal tube to produce a beam. Microwaves are also emitted when certain semiconductors are excited. Water molecules have a natural frequency of 2450 Hz, so will resonate with 2450 Hz microwaves, leading to an increase in KE and hence temperature. Microwaves of this frequency are used in cooking. Since microwaves have a high frequency they can be used to transfer data at the speed required for satellite TV broadcasts and short-range internet links.

Infrared

When a body is given heat, the internal energy of the body increases; in other words the atoms gain energy. Atoms can lose this energy in the form of electromagnetic radiation. The frequency of the radiation depends on the temperature of the body. Bodies at room temperature give out radiation at around 10^{13} Hz. This is classified as infrared. Infrared is used in TV remote controls and optical communications. It is also used in night vision binoculars to see warm objects in the absence of visible light.

Visible light

Visible light is in the range of frequencies that our eyes are sensitive to. Our brains respond to different frequencies by seeing them as different colours: red is the lowest frequency and blue the highest.

Table 1 The visible spectrum.

Colour	Wavelength
violet	380–450 nm
blue	450–495 nm
green	495–570 nm
yellow	570–590 nm
orange	590–620 nm
red	620–750 nm

Ultraviolet

Ultraviolet radiation is produced by high energy electron transitions. Ultraviolet cannot be seen but does cause the emission of visible light from some substances. This is why white clothes glow when illuminated by ultraviolet disco lights.

X-rays

X-ray radiation is high frequency radiation emitted when high energy electrons collide with a metal target. X-rays affect photographic film and can pass through matter. A photograph taken with X-radiation will therefore reveal the inside of an object. This has many applications in medicine.

Gamma radiation (γ)

Gamma radiation is emitted when a nucleus loses energy after a nuclear reaction. These energies are typically in the order of MeVs resulting in radiation with a frequency in the region of 10^{20} Hz. Gamma radiation is even more penetrating than X-rays.

Exercises

Use the spectrum in Figure 12.4 to find out what type of radiation the following wavelengths would be and calculate their frequency:

4 430 nm

5 3.75 m

6 10 μm

7 1 nm

The interaction of EM radiation with matter

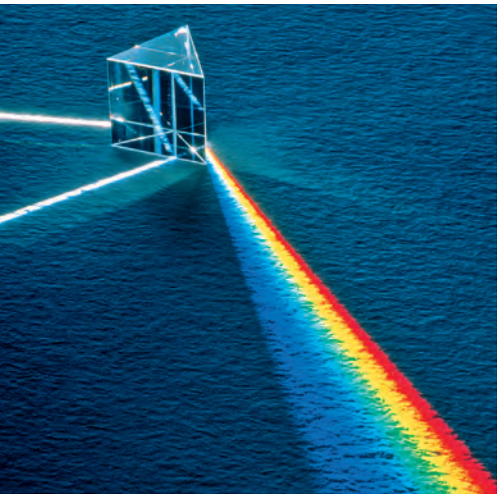
Transmission/absorption

When EM radiation is produced, changing electric and magnetic fields spread out in three dimensions from the source: we say that the wave is transmitted through the medium. The intensity of a wave is the power per unit area, as the wave becomes more spread out its intensity becomes less.

If the power in the whole wave is P then at a distance r this power is spread over a sphere of area $4\pi r^2$. The intensity, I, at a distance, r, is therefore $I = \frac{P}{4\pi r^2}$. In other words $I \propto \frac{1}{r^2}$.
This is called an *inverse square relationship*.

As the wave spreads out it interacts with atoms of the medium.
If an interaction takes place the radiation is absorbed. This can only happen if the energy given up ($\Delta E = hf$) is the correct amount to excite the medium. This is why: microwaves are absorbed by water molecules; IR radiation is absorbed by atoms in solids; and UV radiation is absorbed by the ozone layer. However, the energy in an X-ray is too high to excite an atomic electron and it can pass through most solids.

Radio waves spread out in a sphere centred on the source.



Dispersion of light by a prism.

Reflection

When EM radiation lands on an object it will either be absorbed or transmitted. On absorption the radiation can be re-emitted and this is called reflection. The colour of objects can be explained in terms of reflection and absorption of different wavelengths of light. If a mixture of red, blue and green light is shone onto a blue object the red and green is absorbed but the blue is reflected.

Refraction

EM radiation travels at different speeds in different mediums. When a wave passes from one medium to another the change in speed causes its direction to change. This explains why a ray of light bends when it passes through a block of glass.

Dispersion

The angle of refraction is dependent on the wavelength of the radiation. If red light and blue light both pass into a block of glass the blue light bends more than the red. This is why rainbows are produced when white light passes through a prism.

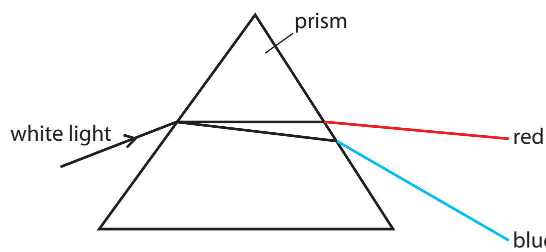


Figure 12.5 When white light is passed through a prism then blue light is refracted more than red light.

Exercises

8 If a light bulb emits 50 W of light what will its intensity be at a distance of 10 m?

9 If intensity of the radiation from the Sun reaching the Earth's atmosphere is $1400\ \mathrm{W\ m^{-2}}$ and the Sun is 146×10^9 m from the Earth, calculate the power of the Sun.

Scattering

When light interacts with small particles such as air molecules or water droplets, it is re-emitted at a different angle, and this causes the light to be scattered. The angle of scattering is dependant on the wavelength of light. Blue light is scattered more than red, this is why the sky and glacier ice look blue.

You can see this glacier because light from the Sun (above) is scattered sideways by the ice. The blue colour is because blue light is scattered most.

EM radiation and health

When electromagnetic radiation is absorbed by human tissue the effect is dependent on the wavelength.

Radio – microwave

When radio waves are absorbed by the body, they cause a slight heating, but do not change the structure of the cells. There seems to be no physical reason why they should cause illness, but cases of illness have been attributed to closeness to a powerful source of radio waves such as a radio antenna. The higher frequency of microwaves used in mobile phone communication means that the heating effects are greater but the power of the signal is weak. There is some evidence that a mobile phone held close to the brain for a long period of time might cause some damage. There is, however, significant risk for people dependent on electronic devices such as pacemakers that interference from strong sources of radio signals can result in malfunction.

It has been reported that the incidence of cancer is higher than normal in a couple of towns that are built close to high power electricity cables. Does this prove that electric fields cause cancer even though the physical principles would suggest not?

IR

The heating effect caused by infrared radiation is significant: exposure to IR can result in burns but low levels of IR cause no harm.

Light

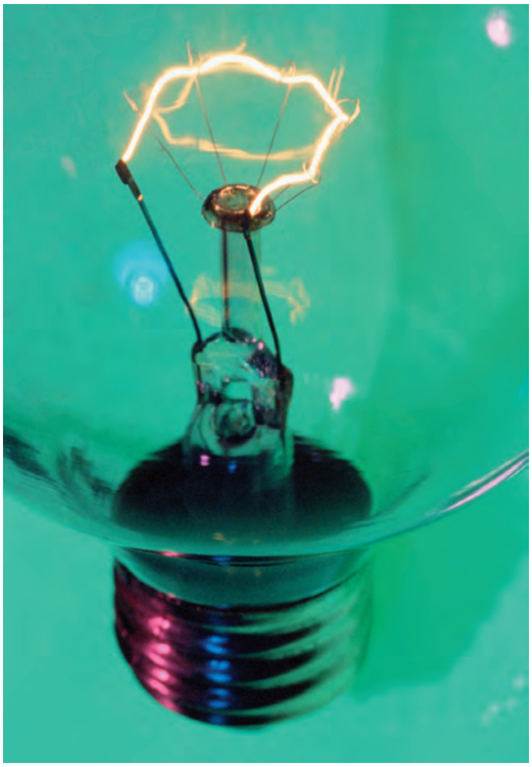
High powered sources of visible light, such as lasers, can damage the eyes and burn the skin.

UV

Exposure to ultraviolet radiation triggers the release of chemicals in the skin that cause redness and swelling. The effect is rather like a burn, hence the name sunburn. UV radiation can also change the structure of the skin's DNA leading to skin cancer.

Ouch!

This site produced by the health protection agency in the UK contains some useful information about radiation and cancer. To access, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 12.2.



The glowing filament of a light bulb.

X-ray and γ ray

Both X-rays and γ (gamma) rays have enough energy to remove electrons from atoms; this is called ionization. When radiation ionizes atoms that are part of a living cell it can affect the ability of the cell to carry out its function or even cause the cell wall to be ruptured. If a large number of cells that are part of a vital organ are affected then this can lead to death. To prevent this there are strict limits to the exposure of individuals to these forms of radiation.

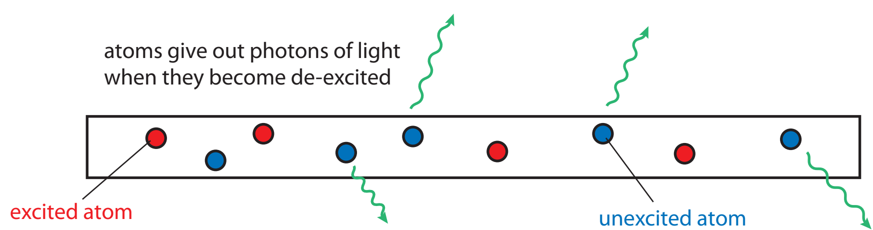
Sources of light

Light is produced when atomic electrons change from a high energy level to a low one. Electrons must first be given energy to reach the high energy level. This can be achieved in a variety of ways.

The light bulb

A light bulb consists of a thin wire filament enclosed in a glass ball. When an electric current flows through the filament, energy is transferred to the filament. This causes the filament to get hot and electrons to become excited (lifted to a higher energy level). Each time an excited electron falls back down to its low energy level a pulse of light is emitted, and these pulses are called *photons*.

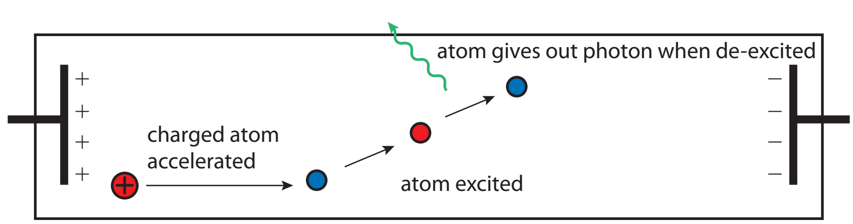
Figure 12.6a An incandescent light bulb



The discharge tube

A discharge tube is a glass tube containing a low pressure gas. A high potential difference created between the ends of the tube causes charged particles in the gas to be accelerated. When these fast moving particles interact with the other gas atoms they excite atomic electrons into high energy levels. When the electrons become de-excited light is emitted.

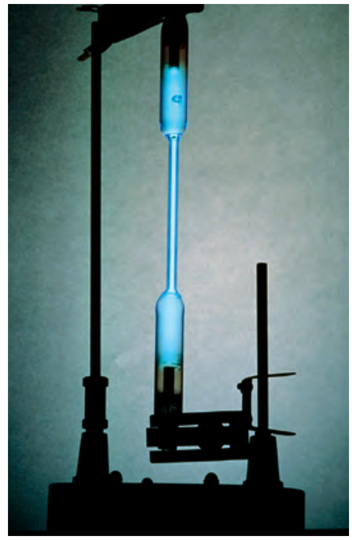
Figure 12.6b A discharge tube

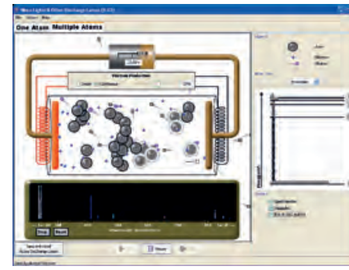


The fluorescent tube

As you can see from the photograph of the discharge tube containing mercury vapour, it does not produce much light. However, a large amount of radiation in the UV region is emitted. This is invisible to the human eye, but if the inside of the tube is coated with a substance that absorbs UV radiation and gives out visible light (fluorescence) then this invisible radiation is converted to light. The result is a much brighter light. This is the principle behind the common strip light, properly called a fluorescent light.

The amount of light from a discharge tube containing mercury vapour is low but the UV radiation is high.





To view this simulation of a discharge tube, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 12.3.

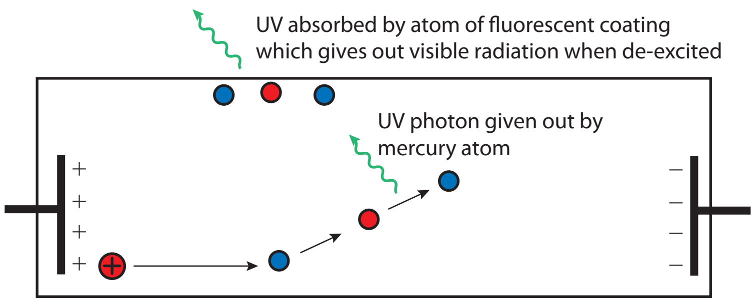


Figure 12.6c A fluorescent tube.

The laser

The laser uses a material with atoms that are able to stay excited for a short time after excitation, e.g. ruby. Electrons are first pumped up to the higher level by a flash of light. This is called population inversion, since there are more excited atoms than non-excited. The excited ruby atoms then start to de-excite, giving out photons of light. This happens in all directions, but some will be emitted along the length of the crystal. These photons will travel past ruby atoms that are still excited causing them to de-excite. The result is an amplification of the light, hence the name LASER (Light Amplification by the Stimulated Emission of Radiation). The amplification can be increased by half silvering the ends so that light is reflected up and down the crystal.

To view this laser simulation, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 12.4.

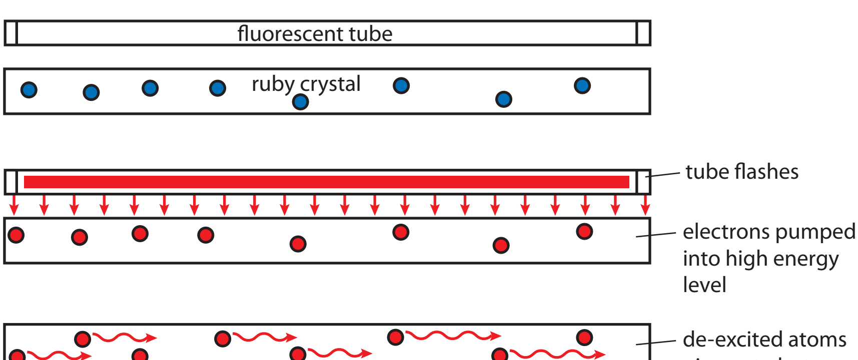


Figure 12.7 The tube flashes, pumping the ruby atoms into the high level. As each photon passes an excited ruby atom it de-excites and another photon is emitted. This is called *stimulated emission of radiation* because the atoms are being stimulated to give out radiation by the passing photon.

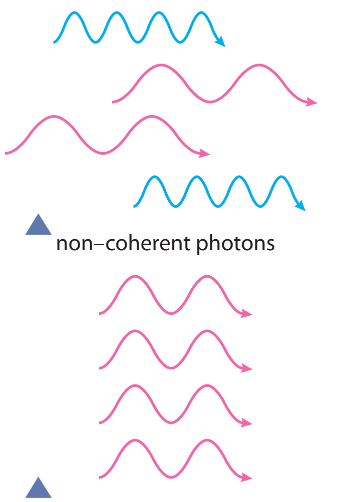


Figure 12.8 The difference between coherent and non-coherent light.

Properties of laser light

Monochromatic

Unlike other light sources, each photon of laser light has the same wavelength; this means the laser is a single colour or *monochromatic*. Light sources giving many wavelengths are white.

Coherence

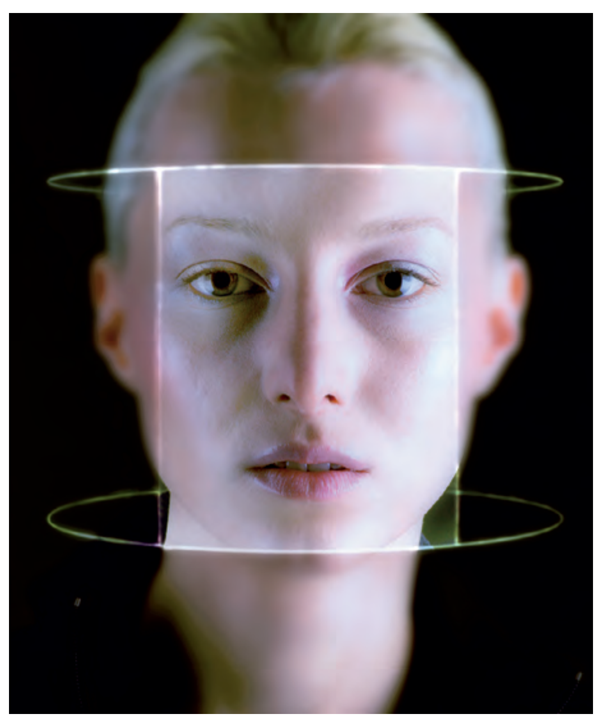
When a *light bulb* emits photons, they are emitted randomly in different directions and with different phase. Each *laser* photon is emitted in the same direction and phase; this is called *coherence*.

Use of lasers

Laser light consists of a parallel beam of coherent light. This means if the beam is split into two parts then, when those two parts are brought together, the interference effects are stable. This property of laser light makes it ideal for the following applications:

- bar code reader
- CD/DVD reader
- production of holograms.

A laser used to scan a face.



The beam can also be made very intense and the fact that it is parallel means that the intensity does not decrease significantly with distance. This makes it possible to use the laser for:

- surgery
- welding
- communications
- measuring devices.

12.2 X-rays

Assessment statements

G.5.1 Outline the experimental arrangement for the production of X-rays.
G.5.2 Draw and annotate a typical X-ray spectrum.
G.5.3 Explain the origins of the features of a characteristic X-ray spectrum.
G.5.4 Solve problems involving accelerating potential difference and minimum wavelength.

Production of X-rays

An X-ray tube is similar to a discharge tube; electrons are accelerated by a potential difference and when they collide with atoms, they excite them, causing them to emit EM radiation. However, X-rays have higher frequency than light waves (therefore higher energy) so the electrons must be accelerated by a higher potential. Also, to have more complete energy transfer, the electrons must collide with the atoms of a solid rather than a gas. The components of an X-ray tube are shown in Figure 12.9.

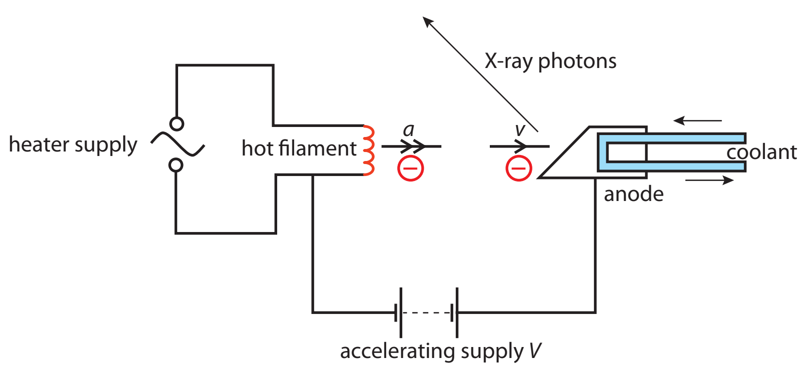


Figure 12.9 An X-ray tube. The electrodes are mounted in a vacuum tube so that the electrons can be accelerated without hitting gas atoms.

Electrons emitted from the hot filament are accelerated towards the anode by a high p.d. When they hit the anode, the KE of the electrons is given to the metal atoms of the anode. This increases the KE of the anode atoms (causing it to get hot) and also excites atomic electrons. Excitation of atomic electrons leads to the emission of photons. If the KE of the electrons is big enough, these photons can be in the X-ray region.

Accelerating potential

X-rays are classified as EM radiation with a wavelength from 10 nm to 0.01 nm. This is equivalent to a frequency from 3×10^{16} Hz to 3×10^{19} Hz. The energy of the lowest frequency X-ray can be found using the equation for the energy of a photon:

$$E = hf = 1.98 \times 10^{-17}\ \text{J (or 124 eV)}$$

According to the law of conservation of energy, if this energy has come from the KE of an electron, the electron must have KE $= 1.98 \times 10^{-17}$ J.

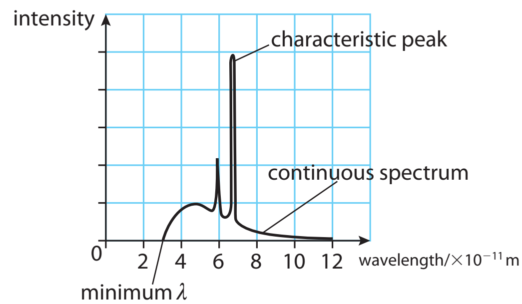
This KE has come from the electrical PE $= Ve = 1.98 \times 10^{-17}$ J

So $V = 124$ V.

The p.d. required to produce a photon of this energy is therefore 124 V. In practice, X-ray tubes operate at potentials of around 50 kV.

The X-ray spectrum

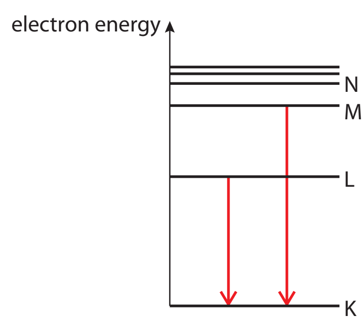
Figure 12.10 The X-ray spectrum for molybdenum from a tube with accelerating p.d. 412 V.



Characteristic peaks

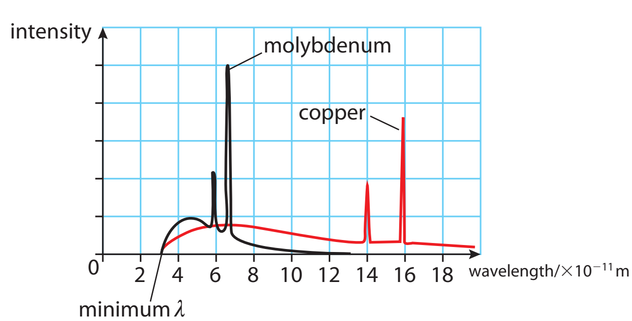
The characteristic peaks in the spectrum are due to emission of photons by atomic electrons in the target material. The frequency of these photons is high, so the atomic electrons must have undergone a big change in energy, such as the ones shown in Figure 12.11. If an electron is excited from the lowest level (K shell) to one of the next two levels, then the photon emitted when that electron goes back down again will be an X-ray photon.

Figure 12.11 Possible energy level changes associated with X-ray emission.



These peaks are different for different target materials, since each element has different electron energy levels.

Figure 12.12 X-ray spectra for molybdenum and copper.



Hardness
The hardness of X-rays is a measure of their penetrating power. Hard X-rays are more penetrating and have a shorter wavelength than soft X-rays.

The continuous spectrum

The rest of the spectrum is continuous; this energy is not emitted by atomic electrons in the target material but by the electrons that hit the target. When these electrons slow down, they emit EM radiation. This is a continuous spectrum, firstly, because all the electrons have different KE and, secondly, not all the energy is converted to X-rays; some KE is given to the target atoms, leading to an increase in temperature (hence the need to cool the target).

The minimum wavelength

The radiation with minimum wavelength corresponds to photons with maximum energy. The photon will have maximum energy if all the KE from the accelerated electrons is converted to photon energy.

$$Ve = \frac{hc}{\lambda_{\text{min}}}$$

We can see from this equation that increasing the accelerating p.d. will result in a decrease in λ_{min} as can be seen in Figure 12.13.

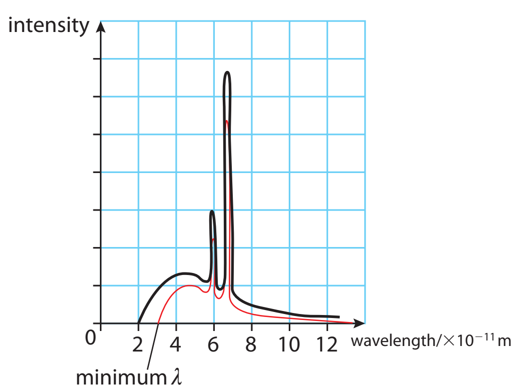


Figure 12.13 Comparing the molybdenum spectrum with two different accelerating p.d.s. The black one has the highest p.d.

Notice that the characteristic peaks are unchanged by changing the accelerating p.d.

Exercises

10 From the minimum wavelength of the X-ray spectrum in Figure 12.10, calculate the accelerating p.d. of the X-ray tube.

11 Estimate the energy in eV of the highest characteristic peak in the molybdenum spectrum. What was the change in electron energy level that gave rise to this line?

12 Figure 12.14 shows two of the electron energy levels for tungsten. What is the wavelength of the photon that would be emitted if an electron went from the higher level to the lower one? Is this an X-ray photon?

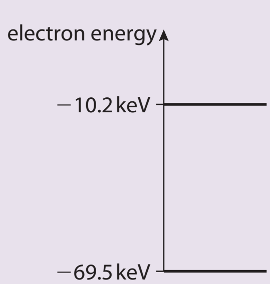


Figure 12.14

12.3 Two-source interference of waves

Assessment statements

- G.3.1 State the conditions necessary to observe interference between two sources.
- G.3.2 Explain, by means of the principle of superposition, the interference pattern produced by waves from two coherent point sources.
- G.3.3 Outline a double-slit experiment for light and draw the intensity distribution of the observed fringe pattern.
- G.3.4 Solve problems involving two-source interference.

Superposition

When two waves of the same type are incident at the same place they add together to give one resultant wave, this is called superposition. The resultant is dependent on the relative phase of the two waves as shown in Figure 12.15.

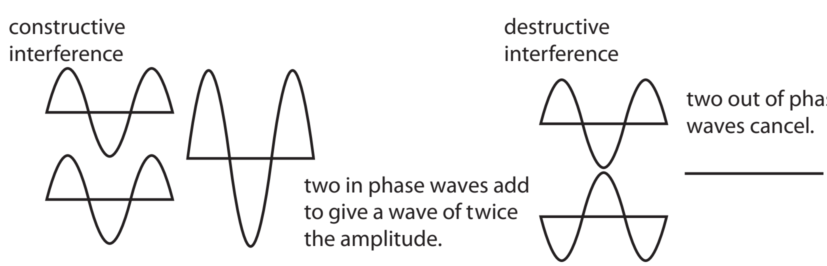


Figure 12.15 Constructive and destructive interference. This effect occurs when two light beams overlap but it can only be observed if the beams are coherent.

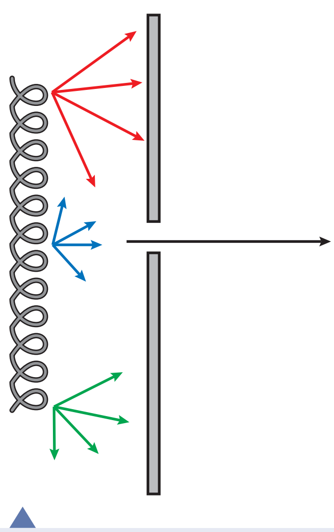


Figure 12.16 A filament bulb is turned into a single source using a narrow slit.

Coherence

Waves that have the same frequency, similar amplitude and constant phase relationship are said to be coherent. As we have seen, the light from a light bulb is emitted randomly so two light bulbs will not be coherent. However, we can make two coherent sources by splitting one source in two, but first we must make one light source as illustrated in Figure 12.16. Different parts of a filament give out light of different wavelength and phase, but using a narrow slit we can select just one part of the filament. This doesn't make the source monochromatic but all parts are in phase.

Note: If laser light is used it is already coherent so the first slit is not necessary.

The double-slit experiment

When the light passes through the narrow slit it spreads out due to diffraction; this makes it possible to pass the light through two more slits.

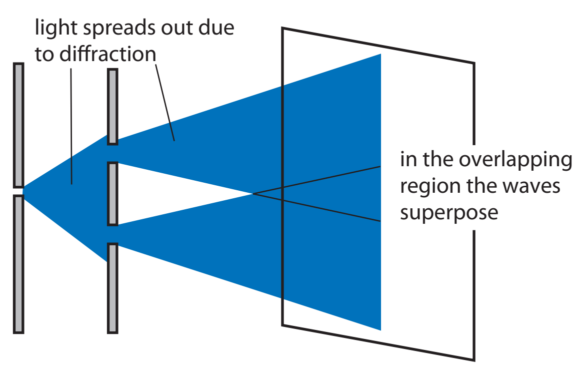


Figure 12.17 Light passing through two narrow slits overlaps due to diffraction.

Phase difference and path difference

Figure 12.18 shows how two waves starting a journey in phase will remain in phase as if they travel the same distance. However, if one wave travels further than the other, they may no longer be in phase.

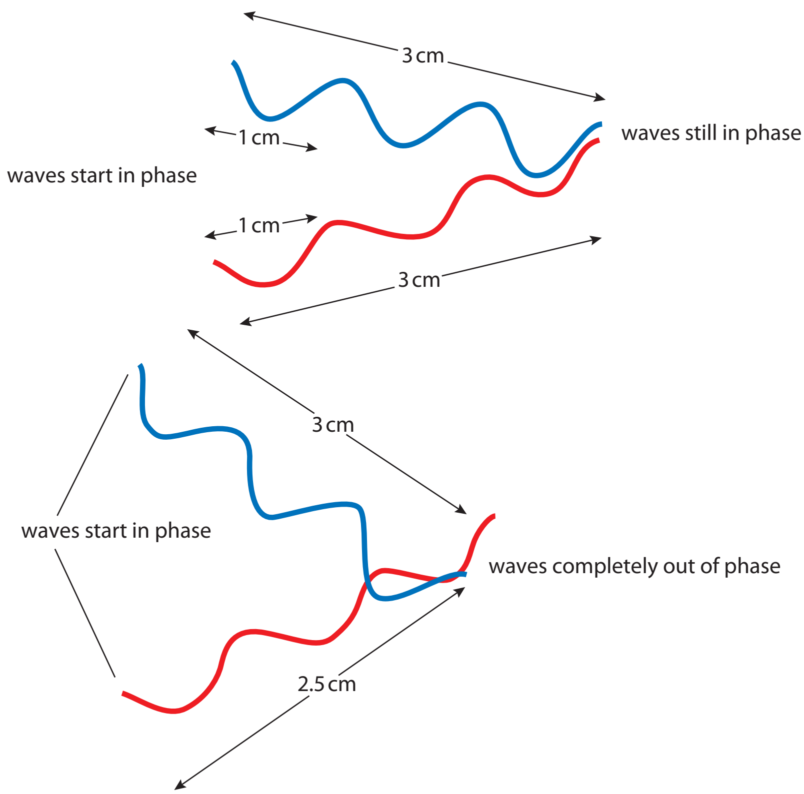


Figure 12.18 Two waves with no path difference and two waves with a path difference of $\frac{1}{2}$ wavelength. In the latter, the blue wave has travelled an extra 0.5 cm, this is the same as $\frac{1}{2}$ a wavelength so at the meeting point a blue peak meets a red trough resulting in destructive interference. This will also happen if the path difference is $1\frac{1}{2}\lambda$, $2\frac{1}{2}\lambda$, $3\frac{1}{2}\lambda$ etc. If the path difference is a whole number of wavelengths then the interference is constructive.

Phase angle ϕ
One whole cycle is equivalent to 2π radians. To convert path difference to phase difference, use the equation

$$\phi = \frac{d}{\lambda} \times 2\pi$$

Exercises

13 Two sources of radio waves separated by a distance of 3 km produce coherent waves of wavelength l00 m. As you walk along a straight line from one station to the other, the signal on your radio is sometimes strong and sometimes weak. This is caused by interference. Calculate whether the signal is strong or weak after walking:

(a) 100 m
(b) 125 m
(c) 250 m.

14 X and Y in Figure 12.19 are coherent sources of 2 cm waves. Will they interfere constructively or destructively at:

(a) A
(b) B
(c) C?

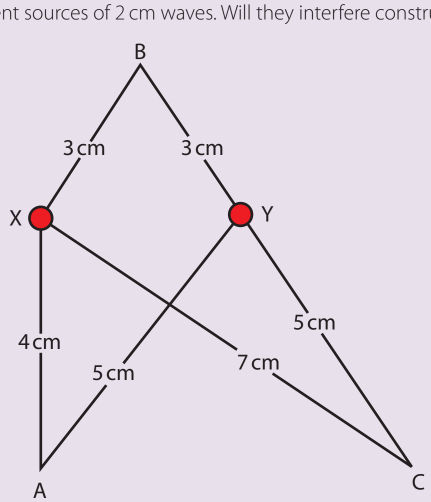


Figure 12.19

Interference of waves from two point sources

When waves from two point sources interfere, the path difference is different at different places; this causes an interference pattern consisting of regions where the waves add and other areas where they cancel.

In Figure 12.20, the 1 cm wavelength waves are interfering and at point 0 waves can be seen. This means constructive interference is taking place. Both waves have travelled a distance of 5 cm. At point X there are no waves, so destructive interference is taking place. However at point C, there are waves again, so there is constructive interference. The path difference is now one complete wavelength.

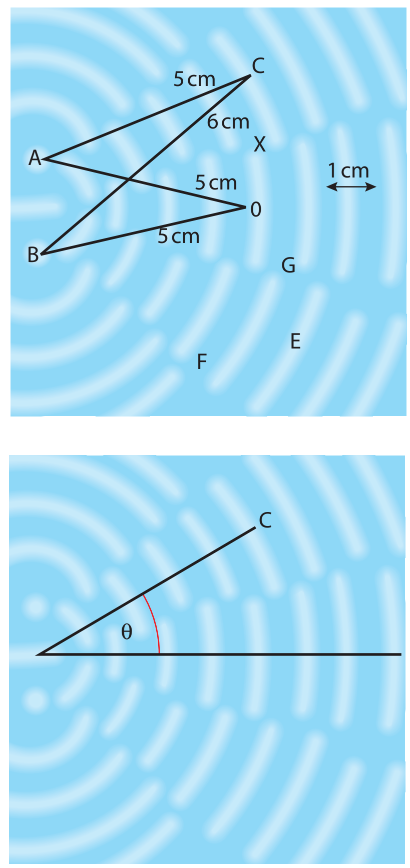


Figure 12.20 The interference pattern from two dripping taps and the radial lines showing the interference effect.

Geometrical model

The position of the different interference effects can be measured using the angle θ that the radial line makes to the middle of the sources as shown in Figures 12.20 and 12.21.

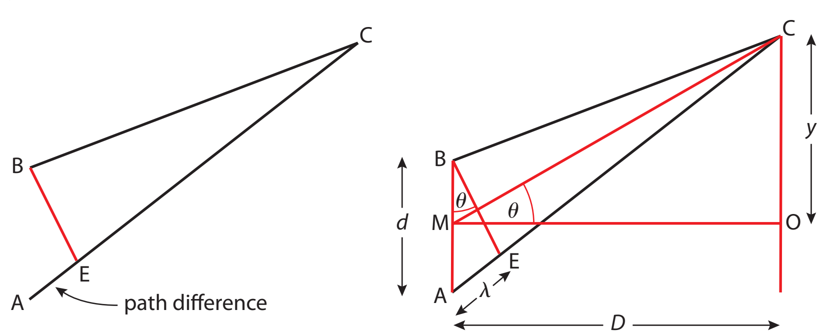


Figure 12.21 Considering point C in Figure 12.20, the path difference is found by drawing the line B-E. This splits the two paths into equal lengths B-C and E-C, the bit left over (A-E) is the path difference. Calculations for two-slit interference can be made with additional construction lines.

C is the position where the path difference is 1λ.

So AE $= \lambda$

If we use light then λ is very small, so the angle θ will be small also.

The angle AEB is almost 90° so $\sin\theta = \frac{\lambda}{d}$

From triangle MCO we see that $\tan\theta = \frac{y}{D}$

Since θ is very small,

$\tan\theta = \sin\theta$, so

$\sin\theta = \frac{y}{D}$

Therefore $\frac{\lambda}{d} = \frac{y}{D}$

The spacing of the interference bands $y = \frac{\lambda D}{d}$

In light, this distance is called the *fringe spacing*.

From this equation we can see that if you make d smaller then y gets bigger. The effect is therefore more visible with sources that are close together.

Exercises

15 Referring to Figure 12.20, what is the path difference at the following points:
(a) E
(b) F
(c) G?

16 Referring to Figure 12.20, two taps separated by 5 cm are dripping into a square tank of water creating waves of wavelength 1.5 cm. The distance to the far side is 1.5 m (D). How far apart will the positions of constructive interference be (y) if measured on the far side of the tank?

17 Calculate y if the taps are moved together so they are now 4 cm apart.

Two-slit interference with light

When two coherent sources of light interfere, destructive interference results in a dark region; constructive interference gives a bright region. Unlike water waves, you can't see light waves interfering as they travel but you can see the effect when they land on a screen. Figure 12.22 shows an example of an interference pattern caused when light passes through two slits. The pattern is a faint series of dots. They are so faint because for diffraction to take place at the slits they must be very narrow.

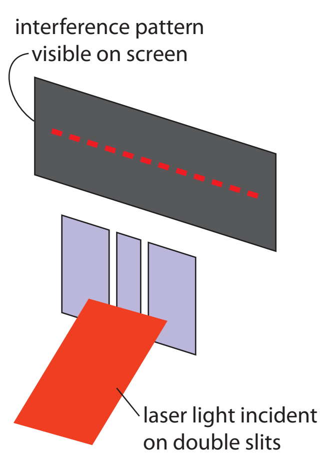


Figure 12.22 An interference pattern

Exercises

18 Two narrow slits 0.01 mm apart (d) are illuminated by a laser of wavelength 600 nm. Calculate the fringe spacing (y) on a screen 1.5 m (D) from the slits.

19 Calculate the fringe spacing if the laser is replaced by one of wavelength 400 nm.

Graphical representation

Using a light sensor it is possible to measure the intensity of light across a diffraction pattern. In this way we can produce a graph of intensity against position as shown in Figure 12.23.

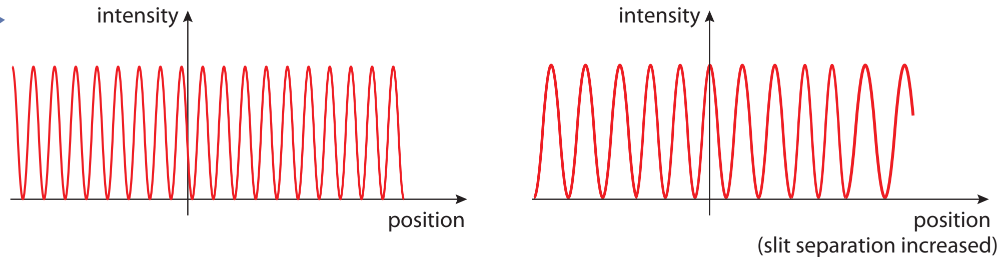


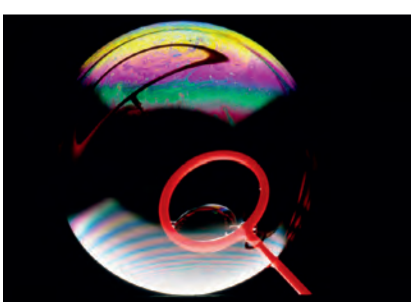
Figure 12.23 Graphs to show intensity against position for double slits. If the slit separation is increased the pattern will spread out .

12.4 Interference by thin films

Assessment statements

G.6.5 State the condition for light to undergo either a phase change of π, or no phase change, on reflection from an interface.
G.6.6 Describe how a source of light gives rise to an interference pattern when the light is reflected at both surfaces of a parallel film.
G.6.7 State the conditions for constructive and destructive interference.
G.6.8 Explain the formation of coloured fringes when white light is reflected from thin films, such as oil and soap films.
G.6.9 Describe the difference between fringes formed by a parallel film and a wedge film.
G.6.10 Describe applications of parallel thin films.
G.6.11 Solve problems involving parallel films.

Colours due to interference of light reflected off a soap bubble.



The condition for observable interference between two sources of light is that the sources are coherent. This means similar amplitude, same wavelength and a constant phase difference. We have seen how this can be achieved using a double slit to split one source in two (division of wavefront). A single source can also be split in two by reflecting half of it off a semi-reflective surface like a bubble (division of amplitude). This results in the coloured bands that we see on the surface of a soap bubble.

Reflection of light off thin films

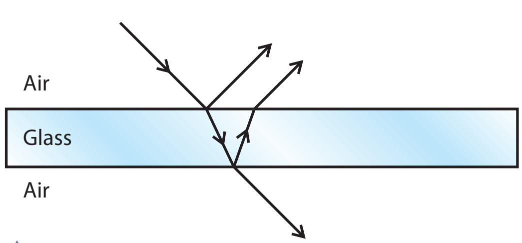


Figure 12.24

When light is incident on a boundary between two different media, e.g. air and glass, part of the light reflects and part refracts. The percentage of light that reflects depends on the media; for glass in air, about 4% of the light is reflected and 96% refracted. If light is incident on a sheet of glass, 4% of the light is reflected off the front surface and then 4% of the remaining 96% off the bottom, as shown in Figure 12.24. If the glass is very thin (about 500 nm) then the two reflected waves will have about the same amplitude, so will interfere if their paths cross. A sheet of glass is too thick for this effect to be seen – however, it can be seen in soap bubbles and oil floating on water.

Phase change on reflection

Before we start to derive a mathematical expression for this effect we should look more closely at what happens when a wave reflects off different surfaces. We can't see what happens when a light wave reflects, so we will consider a wave in a rope as shown in Figure 12.25. A wave in a rope will reflect when it gets to a change in medium. The most extreme cases of this would be if rope is clamped or if it is free to flap about. When the wave peak is incident on a clamped end, the wave tries to push the clamp up. According to Newton's third law, the clamp must exert an equal and opposite force on the rope. This sends a wave trough back down the rope, the wave having undergone a phase change of π. If the end isn't clamped, the reflected wave is the same as the incident wave, so there is no phase change. To visualize how a non-clamped end sends a reflected wave back along the rope, imagine that instead of the end flapping up and down freely, it was your hand moving the rope up and down; if you did this then you would send a wave along the rope.

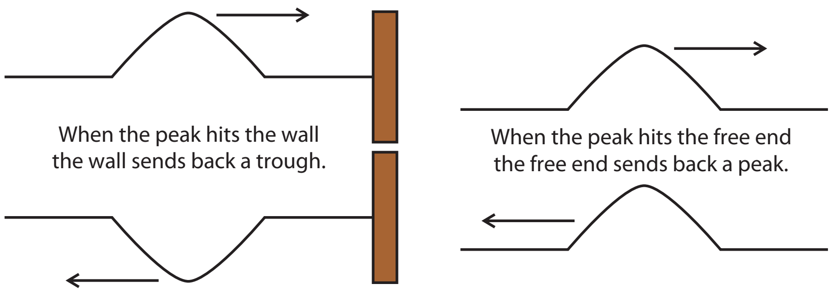


Figure 12.25 Waves in a rope hitting a clamped end and a free end.

To try a simulation of sending a wave along a rope, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 12.5.

If we now relate this to the light we can deduce that when the light reflects off a denser medium (air to glass) then there is a π phase change, but when the light reflects off a less dense medium (glass to air) then there is no phase change.

Interference by parallel-sided thin films

We have seen how light reflects off both surfaces of a thin film producing two coherent sources. To simplify the geometry, we will consider light that is almost perpendicular to the surface of a soap bubble as in Figure 12.26. In this case the parallel reflected rays will coincide when they are focused by the eye.

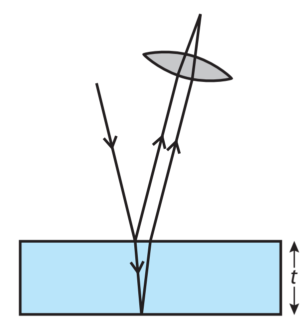


Figure 12.26 Reflected rays brought to a focus by the eye lens.

We can see from this diagram that the path difference is $2t$. This will cause a phase difference of $\frac{2t}{\lambda} \times 2\pi$

But there will also be a phase change of π when the light reflects off the top surface, so the total phase difference $= \frac{4t\pi}{\lambda} - \pi$

Constructive interference will take place if the phase difference is $m2\pi$ where m is an integer (0,1,2…).

So for constructive interference, $\frac{4t\pi}{\lambda} - \pi = m2\pi$

This will take place when $2t = (m + \frac{1}{2})\lambda$

Since the light is travelling in soap, we need to find the wavelength of light in soap. We know that when light passes into soap, it slows down, causing the wavelength to change. The ratio of the speed in air to the speed in soap is given by the refractive index of the air–soap boundary:

$$n = \frac{c_{\text{air}}}{c_{\text{soap}}}$$

Non-perpendicular light

If the light is not perpendicular to the surface, the formula is:

$2nt \cos \varphi = (m + \frac{1}{2})\lambda$ for constructive interference, and

$2nt \cos \varphi = m\lambda$ for destructive interference

where φ is the angle of refraction in the film.

Now $$c = f\lambda$$

so $$n = \frac{f\lambda_{air}}{f\lambda_{soap}} = \frac{\lambda_{air}}{\lambda_{soap}}$$

This gives $$\lambda_{soap} = \frac{\lambda_{air}}{n}$$

Substituting into the equation for constructive interference gives:

$$2t = \frac{(m + \frac{1}{2})\lambda_{air}}{n}$$

And for destructive interference: $2t = m\frac{\lambda_{air}}{n}$

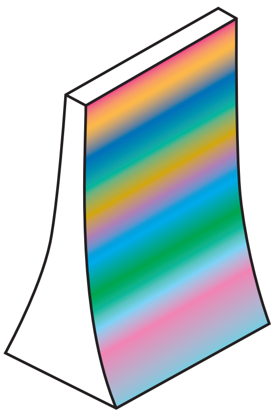


Figure 12.27 The colours in a wedge shaped film.

Let us say that for a certain soap film, the thickness is such that green light ($\lambda = 500$ nm) satisfies this equation. If green light is reflected off the film it will interfere constructively, so the reflected light will be bright – the film looks shiny. If red light is reflected, the interference is destructive, so the film appears dull. If white light, which consists of all wavelengths, is reflected, then only the green will interfere constructively, and the film will therefore appear green.

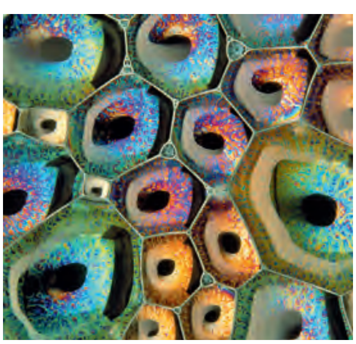
Non-parallel films

Soap bubbles are rarely parallel so different parts of the bubble have different thicknesses. In this case, the wavelength that satisfies the condition for constructive interference will be different in different places. This is why a soap bubble is covered by different colours; the colours can be thought of as contours of thickness. If a soap film is held vertically in a wire frame then the soap will flow towards the bottom, forming a wedge as in Figure 12.27. When viewed in white light, coloured bands can be seen, each band corresponding to a different thickness.

You might notice that the colours on the bands are not the same as rainbow colours. This is because the condition for constructive interference can be satisfied by two different colours at the same place, giving new colours; for example, red and blue gives magenta.

$$(3 + \tfrac{1}{2})\lambda_{blue} = (2 + \tfrac{1}{2})\lambda_{red}$$

Just before the bubbles burst, the top becomes so thin that all wavelengths of light interfere destructively. This can be seen as a dark patch.



Very thin films

If the thickness of the film is very small, the path difference is almost zero, so the only phase difference between the two rays will be π, the phase change of the top ray due to reflection at a more dense medium. This means that all wavelengths will interfere destructively, no light will be reflected, and the film will appear dull as in the photo.

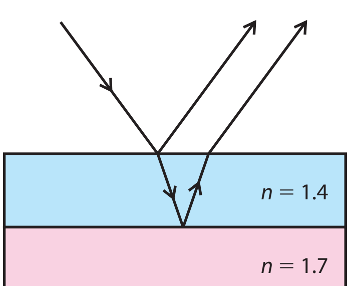


Figure 12.28

Uses of thin films

The main use of thin films is to create anti-reflective coating for lenses and photoelectric cells. Light is a form of energy so must be conserved. If no light reflects off a thin film then the light passing through the film must have increased. Reducing the reflection from lenses therefore makes the image brighter. When light reflects off the upper and lower surfaces of such a coating there will be a phase change of π at both surfaces, so the condition for destructive interference is $t = \frac{\lambda}{4}$. For white light, multi-layer coatings can be applied, one for each wavelength.

Exercises

20 What is the minimum thickness of a soap film that gives constructive interference for light that has wavelength of 600 nm in soap?

21 A camera lens has an antireflective coating that appears violet when viewed in white light. If the wavelength of violet light in the coating is 380 nm, what is the minimum thickness of the coating?

22 Coloured interference fringes are viewed when white light is incident on oil floating on water. The refractive index of oil is 1.5 and the refractive index of water is 1.3.

(a) Will there be a phase change on reflection at the oil/water boundary?

(b) Yellow light has wavelength 580 nm in air. What is its wavelength in oil?

(c) What is the thinnest thickness of oil that will give constructive interference for yellow light?

23 The coating shown in Figure 12.28 is applied to a lens. What thickness should the coating have to remove reflections of light that has wavelength 580 nm in air?

12.5 Interference by air wedges

Assessment statements

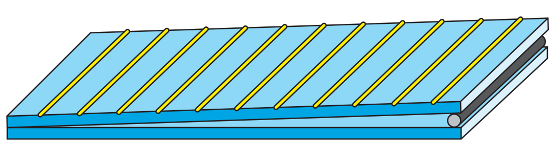
G.6.1 Explain the production of interference fringes by a thin air wedge.

G.6.2 Explain how wedge fringes can be used to measure very small separations.

G.6.3 Describe how thin-film interference is used to test optical flats.

G.6.4 Solve problems involving wedge films.

It is possible to get interference by reflecting light off the two surfaces created when an air wedge is made between two sheets of glass, as shown in Figure 12.29.



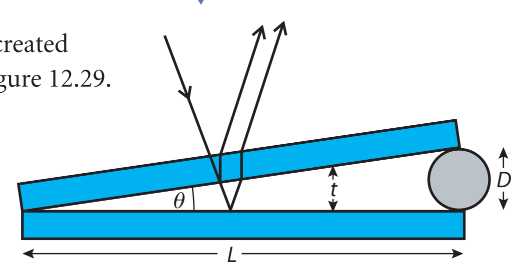


Figure 12.29 Fringes on an air wedge made by placing a hair in between two microscope slides.

Since the thickness of the air wedge increases linearly from one end to the other, the fringes formed are equally spaced straight lines. In common with thin films, a bright fringe will occur when twice the thickness of the wedge, $2t = (m + \frac{1}{2})\lambda$

So the first fringe occurs when $t = \frac{\lambda}{4}$, the second when $t = \frac{3\lambda}{4}$ etc. The difference in wedge thickness between any two adjacent fringes is therefore $\frac{\lambda}{2}$, as shown in Figure 12.30.

The angle θ is very small, so from Figure 12.30 we can see that $\theta = \frac{\lambda}{2x}$ and from Figure 12.29 we can see that $\theta = D/L$. Therefore we can deduce that $\frac{\lambda}{2x} = \frac{D}{L}$

So the fringe spacing $x = \frac{\lambda L}{2D}$

Notice how the thin end of the wedge is dark. This is because the only phase difference is due to the π change of phase on reflection off the lower surface.

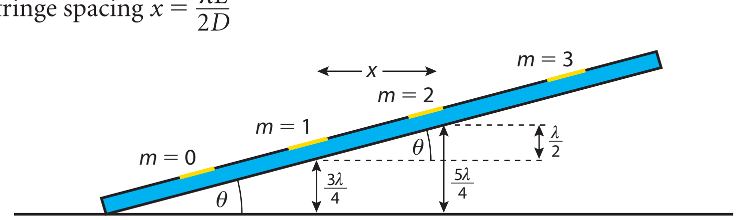


Figure 12.30 The wedge thickness for the 2nd and 3rd fringe.

Application of air wedge interference

The interference fringes produced when a thin object is placed between two glass slides can be used to measure the width of the object. Another useful application of this effect is to test the flatness of a glass surface. If two surfaces are perfectly flat, no fringes will be observed when they are put together. If not perfectly flat, then an air wedge will exist that will produce fringes when illuminated from above. In this way, non-flat areas can be marked and corrected by fine grinding.

Exercises

24 An air wedge is made from two pieces of glass held open by a piece of paper. The wedge is illuminated with light of wavelength 590 nm and the fringes viewed from directly above using a microscope. There are 10 bright lines in a 0.4 cm length.

(a) What is the angle of the wedge?

(b) If the wedge is 5 cm long, how thick is the piece of paper?

(c) If a second piece of paper is added, what will the fringe spacing be?

(d) If water (refractive index 1.3) were put into the wedge with one sheet of paper, what would the new fringe spacing be?

12.6 Diffraction grating

Assessment statements

G.4.1 Describe the effect on the double-slit intensity distribution of increasing the number of slits.

G.4.2 Derive the diffraction grating formula for normal incidence.

G.4.3 Outline the use of a diffraction grating to measure wavelengths.

G.4.4 Solve problems involving a diffraction grating.

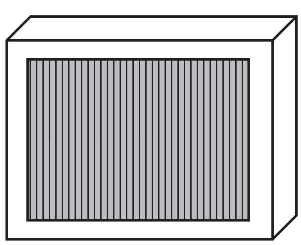


Figure 12.31 Diffraction grating (the number of lines per millimetre can be very high: school versions usually have 600 lines per millimetre).

Multiple-slit diffraction

The intensity of double-slit interference patterns is very low but can be increased by using more than two slits. A diffraction grating is a series of very fine parallel slits mounted on a glass plate.

Diffraction at the slits

When light is incident on the grating it is diffracted at each slit. The slits are very narrow so the diffraction causes the light to propagate as if coming from a point source.

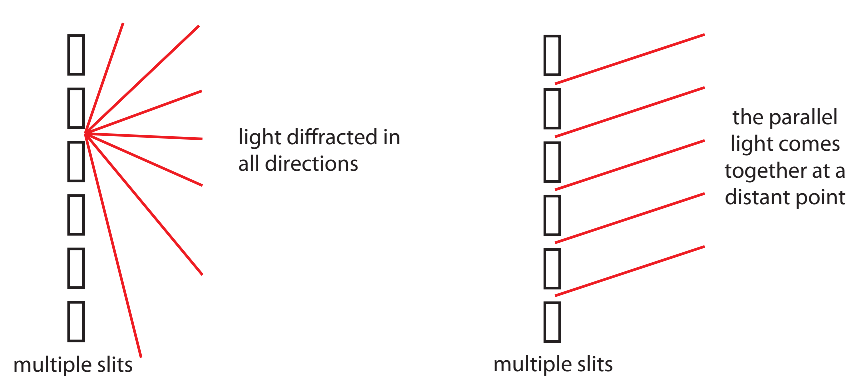
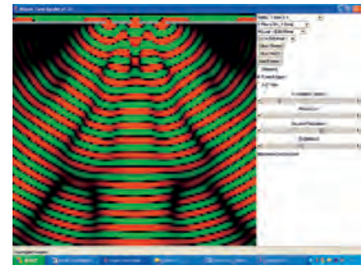


Figure 12.32 Light diffracted at each slit undergoes interference at a distant screen.

Interference between slits

To make the geometry simpler we will consider what would happen if the light passing through the grating were observed from a long distance. This means that we can consider the light rays to be almost parallel. So the parallel light rays diffracted through each slit will come together at a distant point. When they come together they will interfere.



W This simulated pattern shows distinct areas of interference. To view, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 12.6.

Geometrical model

Let us consider waves that have been diffracted at an angle θ as shown in Figure 12.33 (remember light is diffracted at all angles – this is just one angle that we have chosen to consider).

We can see that when these rays meet, the ray from A will have travelled a distance x further than the ray from B. The ray from D has travelled the same distance further than C, and so on. If the path difference between neighbours is λ then they will interfere constructively, if $\frac{1}{2}\lambda$ then the interference will be destructive.

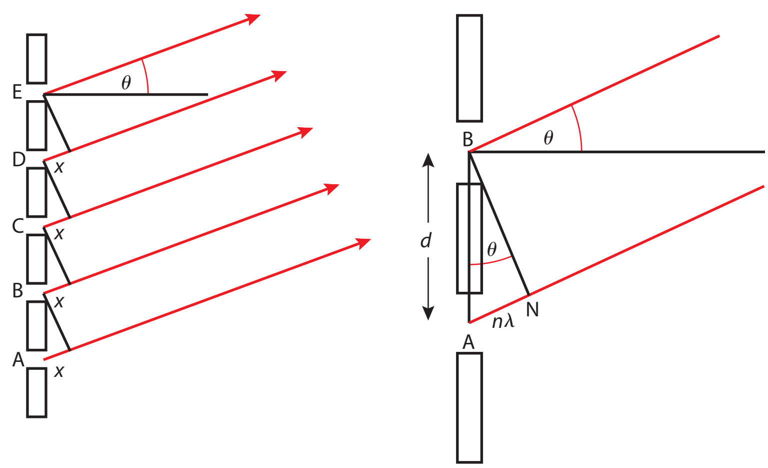


Figure 12.33 Parallel light travels through the grating and some is diffracted at an angle θ. The expansion shows just slits A and B, if the path difference is $n\lambda$ then constructive interference takes place.

The line BN is drawn perpendicular to both rays so angle N is 90°

Therefore from triangle ABN we see that $\sin\theta = \dfrac{n\lambda}{d}$

Rearranging gives $d\sin\theta = n\lambda$

If you look at a light source through a diffraction grating and move your head around, bright lines will be seen every time $\sin\theta = \dfrac{n\lambda}{d}$.

Producing spectra

If white light is viewed through a diffraction grating, each wavelength undergoes constructive interference at different angles. This results in a spectrum. The individual wavelengths can be calculated from the angle using the formula $d\sin\theta = n\lambda$.

Worked example

If blue light of wavelength 450 nm and red light wavelength 700 nm are viewed through a grating with 600 lines mm^{-1}, at what angle will the first bright blue and red lines be seen?

Solution

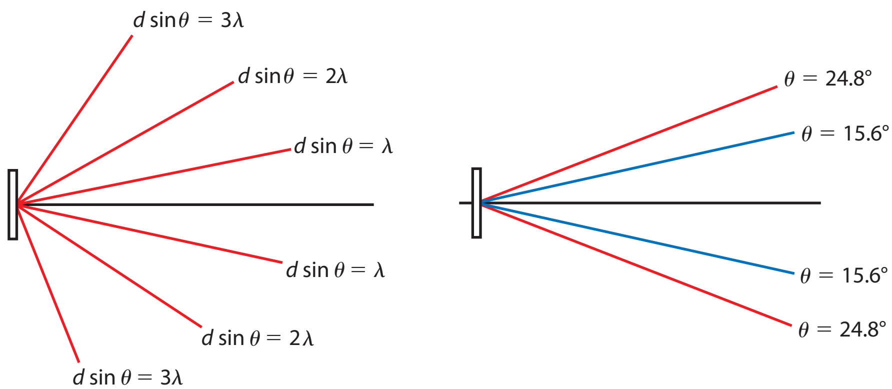


Figure 12.34 Bright lines appear at angles when $\sin\theta = \frac{n\lambda}{d}$. Red and blue lines appear at different angles.

If there are 600 lines/mm, $d = \frac{1}{600}$ mm $= 0.001\,67$ mm

For the first lines, $n = 1$

For blue light, $\sin\theta = \frac{450 \times 10^{-9}}{0.001\,67 \times 10^{-3}} = 0.269$

$\sin\theta = \frac{n\lambda}{d}$

Therefore $\theta_{blue} = \mathbf{15.6°}$

For red light, $\sin\theta = \frac{700 \times 10^{-9}}{0.001\,67 \times 10^{-3}} = 0.419$

Therefore $\theta_{red} = \mathbf{24.8°}$

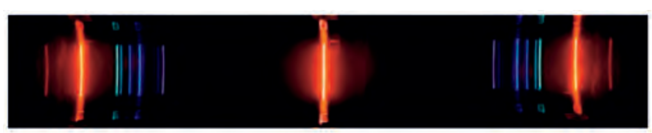


Figure 12.35 A hydrogen lamp viewed through a grating.

Exercises

25 Red light ($\lambda = 700$ nm) is shone through a grating with 300 lines mm^{-1}. Calculate:

(a) the separation of the lines on the grating

(b) the diffraction angle of the first red line.

12.7 X-ray diffraction

Assessment statements

G.5.5 Explain how X-ray diffraction arises from the scattering of X-rays in a crystal.

G.5.6 Derive the Bragg scattering equation.

G.5.7 Outline how cubic crystals may be used to measure the wavelength of X-rays.

G.5.8 Outline how X-rays may be used to determine the structure of crystals.

G.5.9 Solve problems involving the Bragg equation.

We have seen how light is diffracted when it passes through an aperture that is about the same size as its wavelength. However, light is also diffracted when it reflects (see Figure 12.36.) so we could make a diffraction grating out of thin reflective lines. The surface of a CD is made up of very thin lines of aluminium and when light reflects off these lines it diffracts. The interference between light waves from each line gives the coloured bands that you see when you view a CD from different angles. X-rays have a wavelength that is smaller than the size of an atom, so to make a diffraction grating to use with X-rays, the width of the lines would have to be about the size of an atom. We can't make lines that small but we can use crystals.

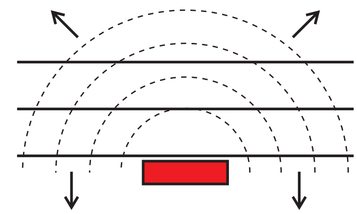


Figure 12.36 A plane wave is diffracted as it reflects off a small object.

Crystals

A crystal has a very regular shape; the reason for this is that the atoms are arranged in a regular way. To make copper sulphate crystals, copper sulphate solution is put in a shallow container and the water allowed to evaporate slowly. In a couple of days the water has gone and you are left with crystals. If you are impatient and try to boil off the water then you just get copper sulphate powder. This is because it takes time for the atoms to settle in the position of lowest PE. There are many different shapes of crystal owing to the different ways that different types of atoms can pack; the simplest is cubic.

Given time, some crystals can grow very big.

Rock crystals
You can determine how quickly a rock cooled down by looking at the crystals it contains. Rocks with big crystals cooled slowly.

X-ray diffraction by crystals

Let us first consider the X-rays diffracted by the top layer of atoms, as in Figure 12.37. Light from the source is incident on the atoms and is diffracted in all directions; this will result in destructive interference unless the path difference is zero, or a whole number of wavelengths.

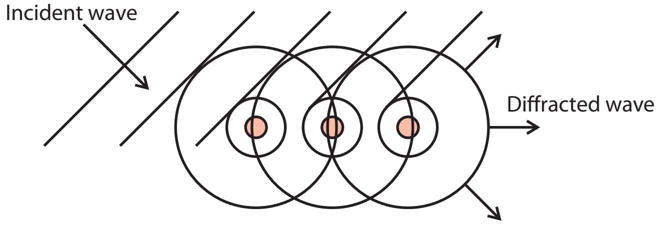


Figure 12.37 Incident wave is diffracted in all directions.

We can see from the construction in Figure 12.38 that if we take the diffracted direction to be the same as the incident ray, then the path difference is zero, which means that the waves will interfere constructively in this direction. In other directions, the X-rays will cancel.

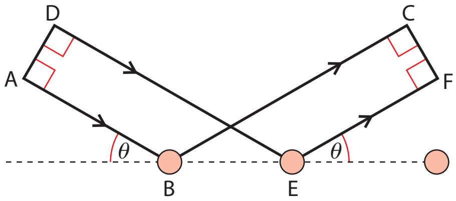


Figure 12.38 When the angle of diffraction equals the angle of incidence, length AB = EF and DE = BC, so the path difference is zero.

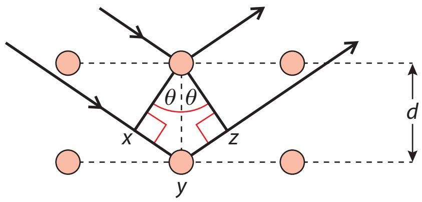


Figure 12.39 X-rays diffracted of the top and second layer.

Let us now consider what happens when X-rays from the top layer interfere with those diffracted by the second layer. As explained above, we will only consider the direction that is the same as the incident ray as in Figure 12.39.

The path difference between these X-rays is now the distance *xyz*. Considering the triangles shown, we can see that $xy = yz = d\sin\theta$ so the path difference $= 2d\sin\theta$

The condition for constructive interference is therefore

$$n\lambda = 2d\sin\theta$$

This is called the Bragg equation.

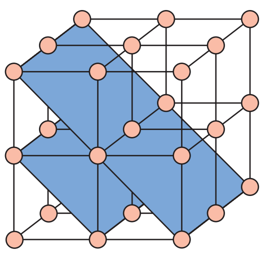


Figure 12.40 Two alternative planes of atoms.

X-ray diffraction patterns

To produce an X-ray diffraction pattern, a narrow beam of X-rays is projected onto a crystal. A photographic plate is then used to record the pattern of diffracted X-rays.

There are many planes of atoms, such as the two shown in Figure 12.40. Each set of planes has different spacing, so will give a different angle for constructive interference. By measuring the different angles, it is possible to determine the arrangement of the atoms.

Single crystals and powders

When a single crystal is used, the diffraction pattern formed is a series of dots. If we were to rotate the crystal, the dots join up to form rings. A powder contains crystals in all orientations so the diffraction pattern will be a set of rings the same as that formed by rotating the crystal.

This X-ray diffraction pattern of DNA by Rosalind Franklin supported Watson and Crick's theory that DNA was a double helix.

Electron diffraction of beryllium.

Exercises

26 In an X-ray diffraction pattern using X-rays of wavelength 5.0×10^{-11} m, the first order ($n = 1$) line is formed at an angle of 10°. Calculate the separation of the atomic planes.

27 A cubic crystal, whose atomic planes are 1.0×10^{-11} m apart, is used to from a diffraction pattern with a first order maximum at 12°. What is the wavelength of the X-radiation used?

28 The first order ring in an X-ray diffraction pattern, using X-rays of wavelength 6×10^{-11} m, has radius of 3 cm. If the distance from the crystal to the photographic plate is 20 cm, calculate the atomic plane spacing.

12.8 Lenses and image formation

Assessment statements

G.2.1 Define the terms *principal axis*, *focal point*, *focal length* and *linear magnification* as applied to a converging (convex) lens.
G.2.2 Define the *power* of a *convex lens* and the *dioptre*.
G.2.3 Define *linear magnification*.
G.2.4 Construct ray diagrams to locate the image formed by a convex lens.
G.2.5 Distinguish between a real image and a virtual image.
G.2.6 Apply the convention 'real is positive, virtual is negative' to the thin lens formula.
G.2.7 Solve problems for a single convex lens using the thin lens formula.

Lenses

A lens is a glass disc that refracts light. If the faces of the disc are curved inwards then the light is caused to spread out (diverge). If the faces curve out then the light is made to focus inwards (converge).

Looking through a convex lens.

Convex lens

Rays of light parallel to the axis converge at the principal focus when they pass through a convex lens as in Figure 12.41 (a).

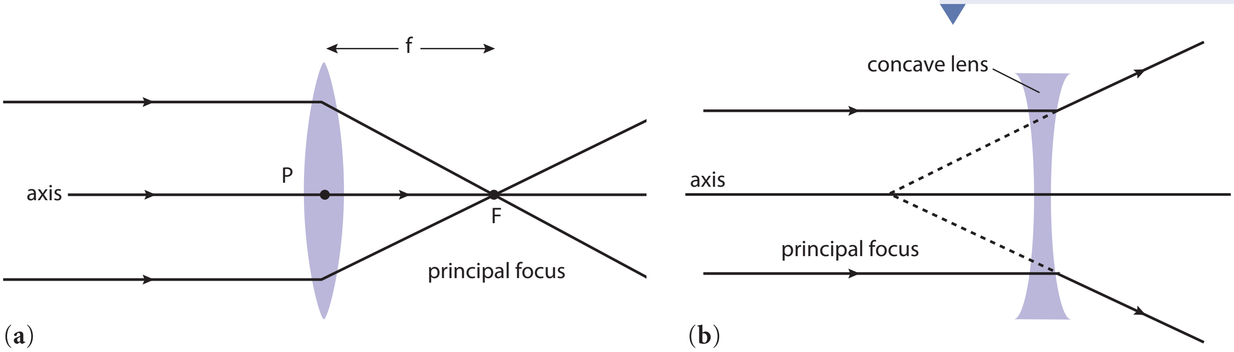
Power of a lens

Lenses with greater curvature bend the light more, resulting in a shorter focal length. Fat lenses are more powerful.

The concave lens

Rays of parallel light are diverged away from the principal focus as in Figure 12.41 (b).

Figure 12.41 The focal length (f) is the distance from the centre of the lens, P (the pole) to the principal focus, F. The optical power of a lens is equal to 1/focal length (units are dioptres).



(**a**) (**b**)

Exercises

29 Parallel light is focused 15 cm from the a convex lens. What is:

(a) the focal length of the lens

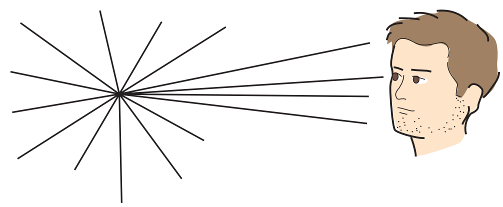
(b) the power of the lens?

Image formation

Point object

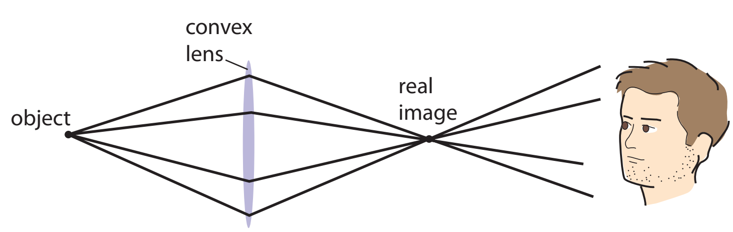
A point object gives out light rays in all directions. When viewed, some of those rays will pass into the eye enabling the observer to see the object. The observer knows that the object is where the rays are coming from. If the light from a point object passes through a lens then the observer will see the light coming from somewhere else, and this is called an *image*. There are two types of image: *real* and *virtual* (see Figure 12.42)

Figure 12.42 Real and virtual images.



Real image

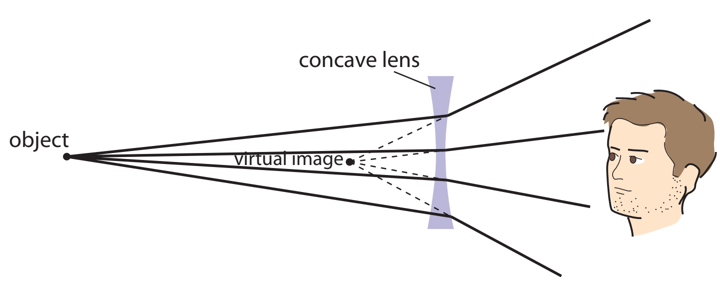
A real image is an image where the rays come from the image. In a convex lens the observer will see the light coming from a point much nearer than the actual object. A real image can be projected onto a screen.



A good example of a virtual image is the image you see when you look into a mirror. It looks like the light is coming from the other side of the mirror but it isn't. This baby doesn't realise that it's just a virtual image.

Virtual image

An image is called a virtual image when the rays only *appear* to come from a point. The rays are coming from the image just the other side of the concave lens.



Extended object

Except when we are looking at stars we rarely look at single point object. Objects normally have size; we call an object with size an extended object. An *extended object* is represented with an arrow and can be treated like two points, one at the top and one at the bottom.

Image formation in convex lenses

The image of an extended object in a convex lens can be found by carefully drawing the path of two rays from the top of the object, the image will be formed where these rays cross or appear to cross.

The nature of the image

We describe the nature of an image according to whether it is:

- real or virtual
- bigger or smaller
- closer or further away
- upright or upside down.

The nature of the image is different for different object positions (see Figures 12.44 and 12.45).

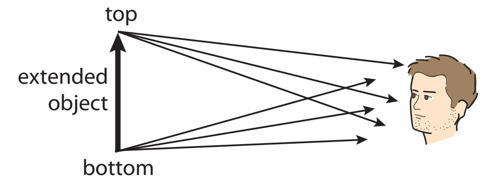


Figure 12.43 An extended object.

Object further than 2 × focal length

How to draw a ray diagram:

- Draw the axis and lens.
- Choose an appropriate scale and mark the principal foci on either side of the lens.
- Measure and draw the object position. This will be given in the question, in this case it is 'more than 2F'.
- Draw a ray from the top of the object parallel to the axis. This ray will be refracted so that it passes through the principal focus (the red ray in Figure 12.44). Refraction takes place at the lens surfaces but for ray diagrams the light can bend at the central line.
- Draw a ray that passes through the centre of the lens. Since the centre of the lens has parallel sides this ray will pass straight through (the blue ray in Figure 12.44).
- The top of the image is the point where the rays cross, and the bottom of the image is on the axis (we don't need to draw rays to find this). Draw the image arrow from the axis to the crossing point.
- The position of the image can now be measured with a rule and scaled up to find the actual image distance.
- The nature of the image can also be determined.

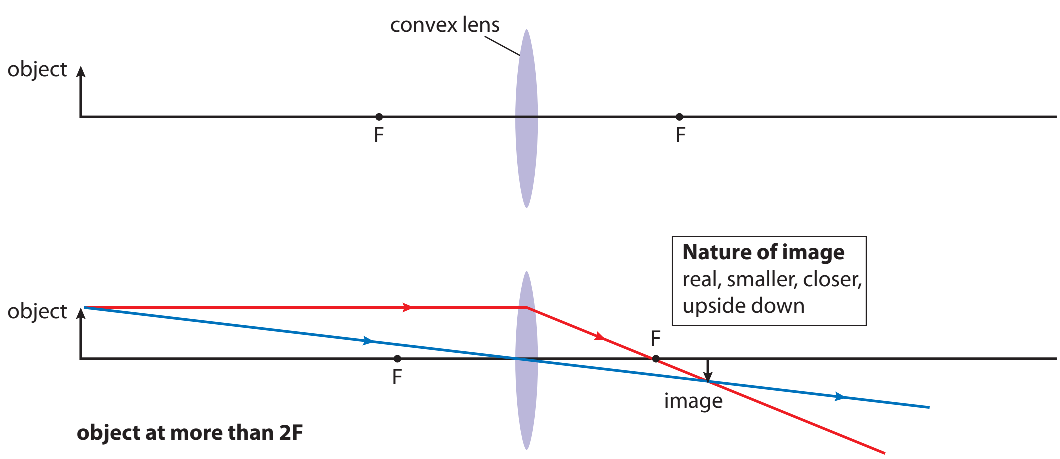
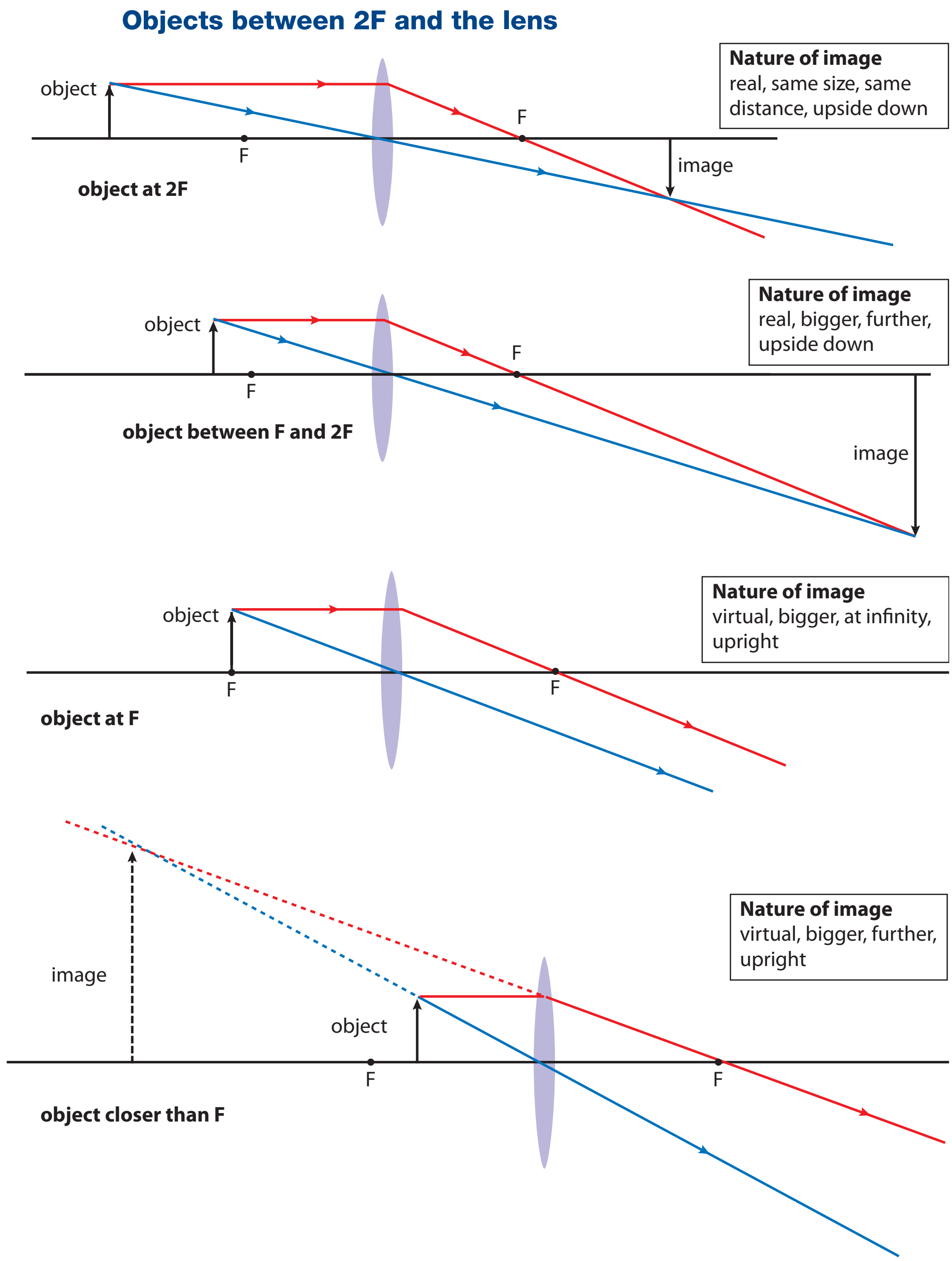


Figure 12.44 Object at more than 2F.

Figure 12.45



● **Examiner's hint:** In this case it looks like there isn't an image but if you were to look at the light coming through the lens your eyes would focus the light like they do when you look at a star. We can say that the light appears to come from infinity.

Linear magnification

Linear magnification is the ratio of image height/object height. For example, if the object is 2 cm high and the image is 6 cm high the magnification is $\frac{6}{2} = 3$

The lens formula

An alternative way of finding the image position is to use the lens formula:

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

f = focal length

u = object distance

v = image distance

A simulation draws ray diagrams for lenses. To try this, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 12.7.

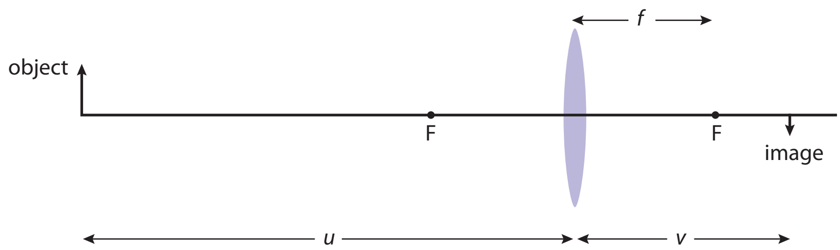


Figure 12.46 Defining the lengths u, v and f.

Worked examples

1 An object is placed 24 cm from a convex lens of focal length 6 cm. Find the image position.

2 An object is placed 3 cm from a convex lens of focal length 6 cm. Find the image position.

Solution

1 From the question:

$u = 24$ cm

$f = 6$ cm

rearranging: $\frac{1}{f} = \frac{1}{u} + \frac{1}{v}$ gives $\frac{1}{v} = \frac{1}{f} - \frac{1}{u}$

substituting values: $\frac{1}{v} = \frac{1}{6} - \frac{1}{24} = \frac{4}{24} - \frac{1}{24} = \frac{3}{24}$

so $v = \frac{24}{3} = 8$ cm

2 From the question:

$u = 3$ cm

$f = 6$ cm

rearranging: $\frac{1}{f} = \frac{1}{u} + \frac{1}{v}$ gives $\frac{1}{v} = \frac{1}{f} - \frac{1}{u}$

substituting values: $\frac{1}{v} = \frac{1}{6} - \frac{1}{3} = \frac{1}{6} - \frac{2}{6} = \frac{-1}{6}$

so $v = \frac{-6}{1} = -6$ cm

● **Examiner's hint:** Always draw a sketch of the relative positions of the object, image and lens. This will help you to see what the problem involves. Then use the formula to find the thing you are asked to calculate.

Sign convention

We know from the ray diagrams that the image in Worked example 1 is real and the one in Worked example 2 is virtual. We can see that the images are different by looking at the sign of the answer.

REAL IS POSITIVE

VIRTUAL IS NEGATIVE

This convention applies to focal lengths and object distances too.

Linear magnification

From the definition the linear magnification of the image in Figure 12.47 is $\frac{h_i}{h_o}$

But we can see that the blue ray makes two triangles with the same angle,

therefore $\frac{h_i}{h_o} = \frac{v}{u}$

Linear magnification, $M = \frac{v}{u}$

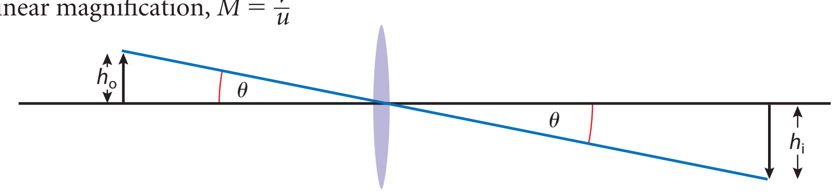


Figure 12.47 Linear magnification.

Exercises

Use the lens formula to solve the following problems. You can check your answers by drawing ray diagrams too.

30 A 25 cm focal length lens is used to focus an image of the Sun onto a piece of paper. What will the distance between the lens and the paper be?

31 An object is placed 30 cm from a convex lens of focal length 10 cm.
- **(a)** Calculate the image distance.
- **(b)** Is the image real or virtual?
- **(c)** Calculate the magnification of the image.

32 A real image is formed 20 cm from a convex lens of focal length 5 cm. Calculate the object distance.

33 An object is placed 5 cm from a lens of focal length 15 cm,
- **(a)** Calculate the image distance.
- **(b)** Is the image real or virtual?
- **(c)** Calculate the magnification of the image.

34 A camera with a single lens of focal length 5 cm is used to take a photograph of a bush 5 m away. A simple camera uses a convex lens.
- **(a)** What is the object distance?
- **(b)** Calculate the distance from the lens to the film (v).
- **(c)** What is the linear magnification of the camera?
- **(d)** If the bush were 1 m high how high will the image be?

35 The camera of Question 34 is used to take a picture of a flower on the bush so the photographer moves towards the bush until he is 20 cm from the flower.
- **(a)** Calculate the image distance.
- **(b)** What is the linear magnification?

12.9 Optical instruments

Assessment statements

G.2.8 Define the terms *far point* and *near point* for the unaided eye.

G.2.9 Define *angular magnification*.

G.2.10 Derive an expression for the angular magnification of a simple magnifying glass for an image formed at the near point and at infinity.

G.2.11 Construct a ray diagram for a compound microscope with final image formed close to the near point of the eye (normal adjustment).

G.2.12 Construct a ray diagram for an astronomical telescope with the final image at infinity (normal adjustment).

G.2.13 State the equation relating angular magnification to the focal lengths of the lenses in an astronomical telescope in normal adjustment.

G.2.14 Solve problems involving the compound microscope and the astronomical telescope.

In this section, we are going to investigate three optical instruments: the magnifying glass, the telescope and the microscope. All three instruments enable us to see an object more clearly, but first we should see how well we can do with the unaided eye.

The human eye

Inside the eye there is a convex lens. This, together with the front part of the eye, focuses light onto the retina, where millions of light sensitive cells sense the light and send electrical signals to the brain.

The eye lens is made of a rubbery substance that can be squashed; squashing the lens makes it fatter and therefore more powerful. In this way the eye can be adapted to focus on objects that are close or far away as illustrated in Figure 12.49. There is a limit to how fat the lens can get. If an object is too close to the eye, then it can't focus the rays on the retina, and the image is 'out of focus'. The average closest distance is 25 cm, but this tends to get longer with age.

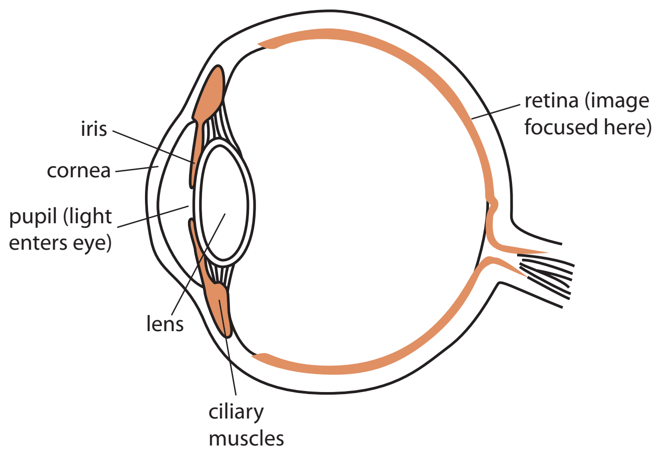


Figure 12.48 Parts of the human eye.

● **Hint:** SL Option A students can find out more about the eye and sight in the Appendix on page 577.

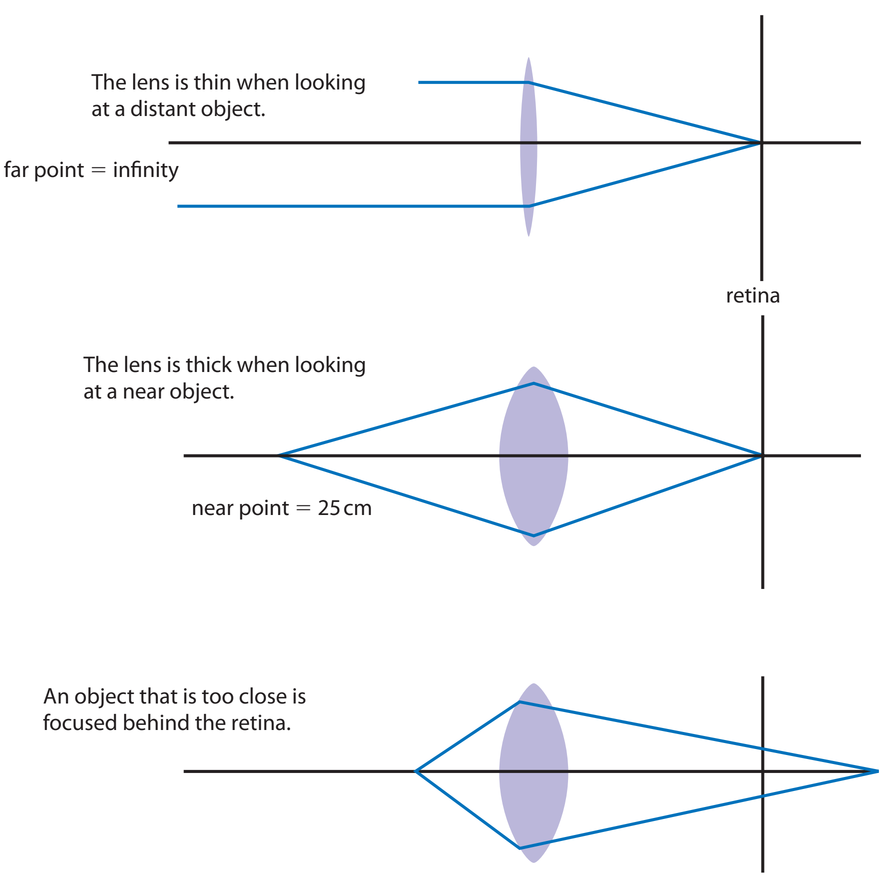


Figure 12.49 The eye lens changes shape to view different objects.

Short sight
Short sighted people have a near point that is closer to the eye than normal. This means they can read things that are much closer. Unfortunately it also means that they cannot see things that are far away.

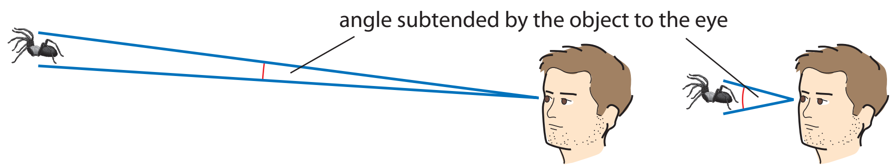
Although the wind turbines are all the same size the nearest one looks bigger.

The size of the Moon
You may have noticed that the Moon looks bigger when it is just above the horizon than it does when it is up above. This is in fact an illusion, if you measure the size of the Moon you find it never changes. Your brain decides how big something is depending on how your eyes are focused. When the Moon is on the horizon your brain thinks it is closer because of the other objects in view. This is an example of how perception sometimes doesn't agree with measurement.

How big does an object appear?

We are all familiar with the fact that objects that are far away seem smaller than objects that are close. We can measure how big something appears using the angle that rays make when they enter the eye. In Figure 12.50 a) we can see how the object subtends a bigger angle when viewed from a short distance. If we want to make an object appear as big as possible then we should view it as near as possible. This means at a distance of 25 cm.

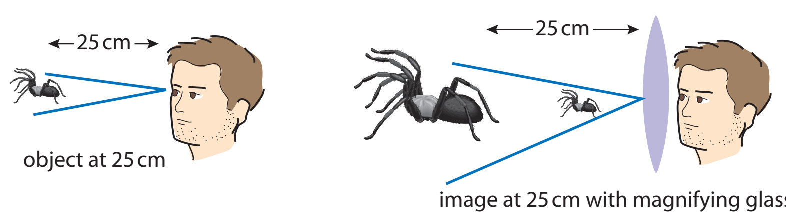
Figure 12.50 a) A close object appears bigger.



The magnifying glass

We use a magnifying glass to make things look bigger; this is done by putting the object closer than the principal focus of a convex lens. Without a magnifying glass the best we can do is to put an object at our near point (25 cm in average eyes). The best we can do with a magnifying glass is with the *image* at the near point.

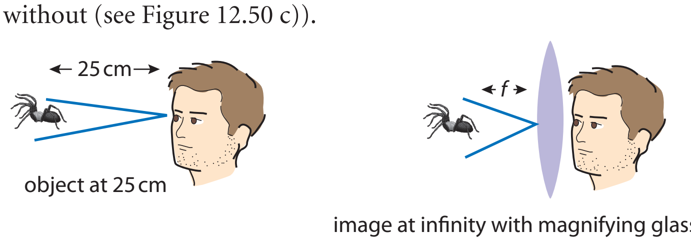
Figure 12.50 b) Using a magnifying glass with the image at 25 cm.



The problem with looking at something so close is that it can be a bit tiring, since your eye muscles have to squash the lens.

It is more relaxing to view the image at a distance, and then the eye is relaxed. This however doesn't give such a magnified image. If the final image is far away (we could say an infinite distance) the rays coming to the observer should be parallel. In the previous section we saw that this means the object must be at the focal point. In both cases the angle subtended when using the magnifying glass is bigger than without (see Figure 12.50 c)).

Figure 12.50 c) Using a magnifying glass with the image 'at infinity'.



Angular magnification (M)

The angular magnification tells us how much bigger an object looks.

$$\text{Angular magnification} = \frac{\text{angle subtended by image at eye } (\beta)}{\text{angle subtended by object at unaided eye } (\alpha)}$$

Angular magnification for a magnifying glass

1. Image at infinity

When the final image is an infinite distance away the object must be placed at the focal point. Looking at Figure 12.51, you can see why this image looks bigger than the image in the unaided eye.

If the angles are small (for the original object with the image at infinity) and measured in radians then

$$\alpha = \frac{h_o}{25}$$

$$\beta = \frac{h_o}{f}$$

Since angular magnification $M = \frac{\beta}{\alpha} = \frac{h_o}{f} \times \frac{25}{h_o}$

So $M = \frac{25}{f}$

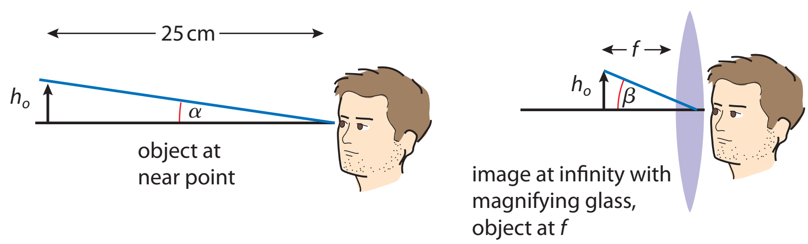


Figure 12.51 a) Angular magnification for an original object with image at infinity and viewed with a magnifying glass.

2. Image at the near point (normal adjustment)

Figure 12.51 b) compares an object as close as possible to the unaided eye to the same object viewed with a magnifying glass. So that the final image is also as close as possible, the object must be placed close to the lens.

This can be shown to give an angular magnification of $1 + \frac{25}{f}$. (One more than the previous example.)

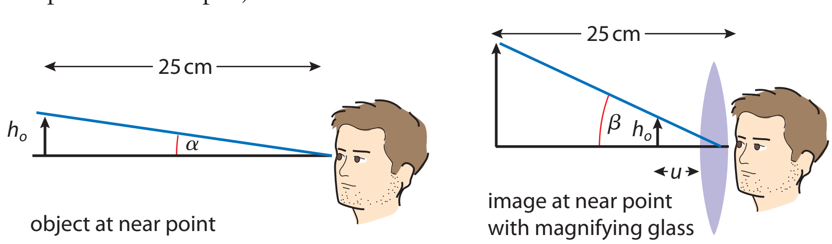


Figure 12.51 b) Angular magnification for an image at the near point.

Derivation of $M = 1 + \frac{25}{f}$

Referring to Figure 12.51 b), if the angles are small then the angles expressed in radians are:

$\alpha = \frac{h_o}{25}$

$\beta = \frac{h_o}{u}$

so $M = \frac{\beta}{\alpha} = \frac{h_o}{u} \times \frac{25}{h_o}$

$M = \frac{25}{u}$ (1)

but $\frac{1}{f} = \frac{1}{u} + \frac{1}{v}$ so $\frac{1}{u} = \frac{1}{f} - \frac{1}{v}$

but $v = -25$ cm so $\frac{1}{u} = \frac{1}{f} + \frac{1}{25}$

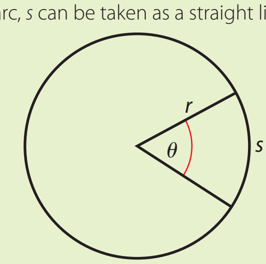
Rearranging gives $u = \frac{25f}{25 + f}$

Substituting for u in equation (1)

gives $M = \left(\frac{25 + f}{f}\right) = 1 + \frac{25}{f}$

So $M = 1 + \frac{25}{f}$

Remember the radian $= \frac{s}{r}$
If the angle is very small then the arc, s can be taken as a straight line.



Exercises

36 The Moon is about 3500 km in diameter and about 400 000 km away from the Earth. Estimate the angle subtended by the Moon to an observer on the Earth.

37 If a small insect 1 mm long is viewed at a distance of 25 cm from the eye, what angle will it subtend to the eye?

38 How close to a lens of focal length 5 cm should the insect of Question 37 be placed so that an image is formed 25 cm from the eye?

39 Use the formula to calculate the angular magnification of the insect viewed with a lens of focal length 5 cm if the final image is at the near point.

● **Examiner's hint:** If you draw the object before you draw the first ray you often end up with a final image that doesn't fit on the page (try it).

However, if the position of the object is given in the exam question you have to use it. The examiner will have made sure that everything will fit okay, so draw the ray in Step 1 from the top of the object and continue through the steps.

The microscope

The microscope is used to produce an enlarged image of a close object. The microscope consists of two convex lenses: the one closest to the object is called the *objective*; and the one you look through is the *eyepiece*. To give maximum magnification, the final image is at the near point of the eye. Figure 12.52 shows the ray diagram for a point object. This is called *normal adjustment*.

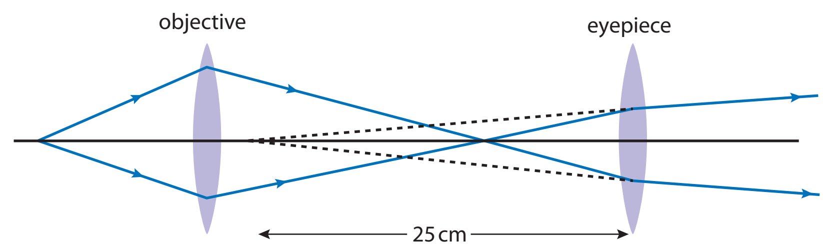


Figure 12.52 Simple ray diagram for a microscope.

Drawing the ray diagram with an extended object

Drawing the ray diagram for an extended object is a bit more difficult, but you need to know how to do it for the exam.

1 Draw the lenses and axis then a ray through the centre of the objective to a point half way down the eyepiece. Then draw an object a short distance from the objective.
2 Draw a ray from the object parallel to the axis. Continue this ray so that it hits the bottom of the eyepiece. Now mark F_o, it is the point where this ray crosses the axis.
3 To find the position of the final image draw a construction line (black) from the top of the first image through the middle of the eyepiece. The top of the image will lie on this line. Choose a point on this line beyond the objective and draw the rays coming from this point. Now add arrows to all the rays.

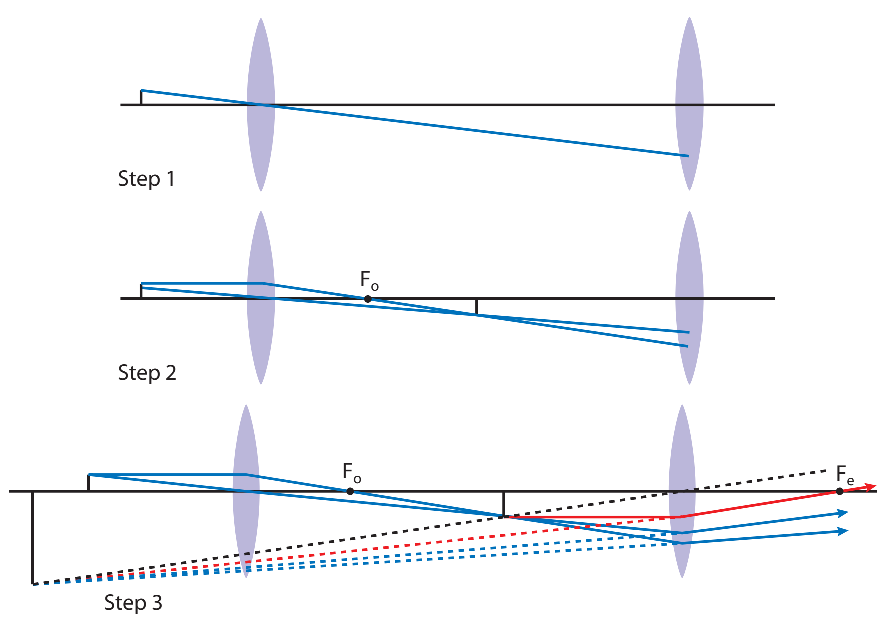


Figure 12.53 The steps in drawing a ray diagram with an extended object.

To find the focal point of the eyepiece, the red construction line can be drawn. This comes from the top of the first image and goes parallel to the axis. When it passes through the lens it appears to come from the top of the final image. The ray will pass through the focal point.

The astronomical telescope

The astronomical telescope is used to view stars and planets. It has no use for looking at things on Earth since the image is upside down. The simple telescope consists of two convex lenses that are used to produce a virtual image of a distant object at infinity (see Figure 12.54) The final image could be produced anywhere but in normal adjustment it is at infinity, in other words the rays come out parallel.

The objective lens forms an image of a distant object at its principal focus. This image is at the principal focus of the eyepiece, so the final image is at infinity (the rays are parallel).

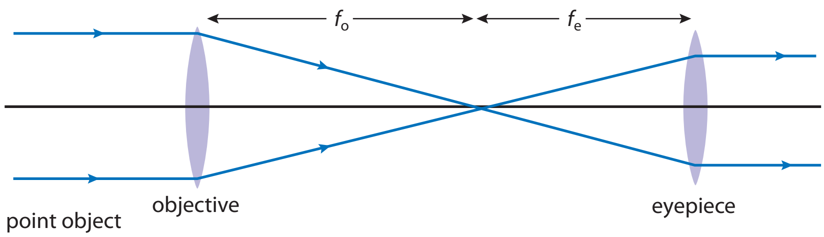


Figure 12.54 a) The astronomical telescope (point object).

To see how this produces a magnified image we must use an extended object. In the case in Figure 12.54 b) the object is the Moon. The three blue rays are coming from the top of the Moon and the bottom of the Moon is in line with the axis. By the time the rays reach the Earth they are very nearly parallel.

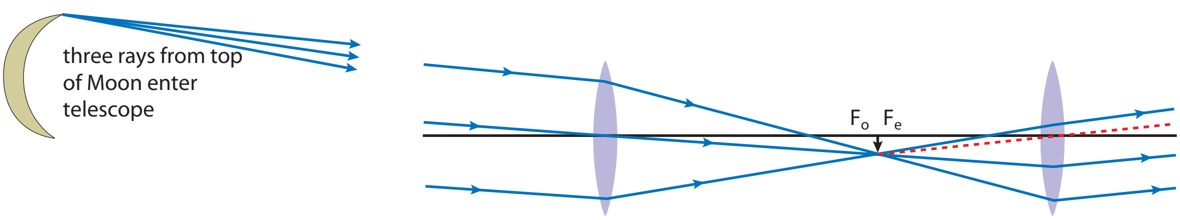


Figure 12.54 b) The astronomical telescope (extended object).

How to draw the ray diagram

This diagram looks difficult to draw but is okay if done in stages.

1 Draw the lenses and axis but don't draw the foci yet.

2 Draw a ray passing through the centre of the objective hitting the eyepiece about half way down.

3 Draw two more rays entering the objective at the same angle as the first. Then draw the top ray hitting the bottom of the eyepiece.

4 The bottom ray will cross the other two at the same place; this is just below the principal focus. You can now mark this on the axis and draw in the first image (F_o).

5 The rays emerge from the eyepiece parallel. To find the angle, draw a construction line (dotted) from the top of the image straight through the centre of the eyepiece. All the rays will be parallel to this. Add arrows to all the rays.

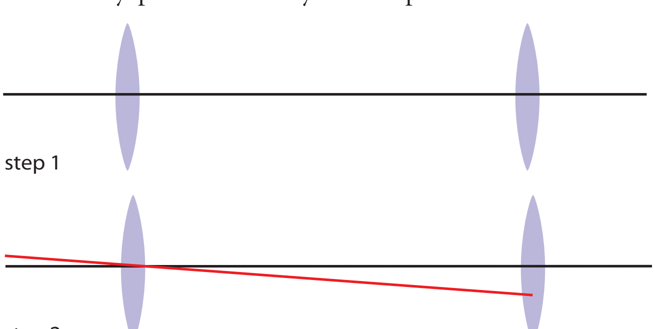
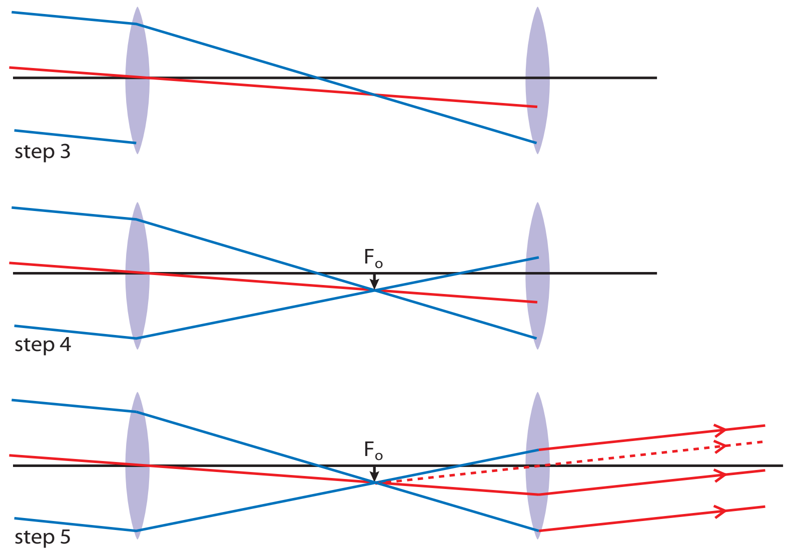


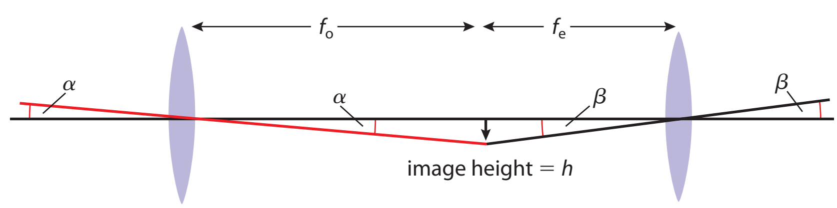
Figure 12.55 Steps in drawing the ray diagram for an astronomical telescope.



Angular magnification

Since the object is very far away, the image subtended to the unaided eye is the same as the image subtended to the telescope. The angles subtended by the object and the image are as shown in Figure 12.56.

Figure 12.56 Angles subtended at the lenses in an astronomical telescope.



A telescope with a large angular magnification is very long. The problem then is that not much light can travel through it; this means that lenses with a large diameter should be used. These are difficult to make, so most high powered telescopes use mirrors not lenses.

If the angles are small and measured in radians then:

$$\propto = \frac{h}{f_o}$$

$$\beta = \frac{h}{f_e}$$

$$\text{angular magnification} = \frac{\beta}{\propto} = \frac{h}{f_e} \times \frac{f_o}{h} = \frac{f_o}{f_e}$$

● **Examiner's hint:** Before starting the calculation, draw a sketch showing the different positions of the lenses and object. Don't try to draw the rays, the sketch is just to help you see the relative positions.

Exercises

40 A microscope is constructed from an objective of focal length 1 cm and an eyepiece of focal length 5 cm. An object is placed 1.5 cm from the objective.

(a) Calculate the distance from the objective to the first image.

(b) If the final image is a virtual image 25 cm from the eyepiece, calculate the distance between the first image and the eyepiece.

(c) Calculate the distance between the lenses.

41 A telescope is constructed from two lenses: an objective of focal length 100 cm and an eyepiece of focal length 10 cm. The telescope is used in normal adjustment (final image at infinity):

(a) Calculate the angular magnification.

(b) What is the distance between the lenses?

42 A telescope has an objective of focal length 50 cm. What focal length eyepiece should be used to give a magnification of 10?

To access a virtual optics lab, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 12.8.

12.10 Aberrations

Assessment statements

G.2.15 Explain the meaning of spherical aberration and of chromatic aberration as produced by a single lens.

G.2.16 Describe how spherical aberration in a lens may be reduced.

G.2.17 Describe how chromatic aberration in a lens may be reduced.

We have assumed in all the previous examples that parallel rays of light are brought to a point when they shine through a convex lens. However, this is not the case with a real lens.

Spherical aberration

Because of the spherical curvature of a lens, the rays hitting the outer part are deviated more than the ones on the inside (see Figure 12.57).

The result is that if the image is projected onto a screen there will be a spot instead of a point. If such a lens were used to take a photograph then the picture would be blurred. To reduce this effect, the outer rays are removed by placing a card with a hole in it over the lens. This is called *stopping*.

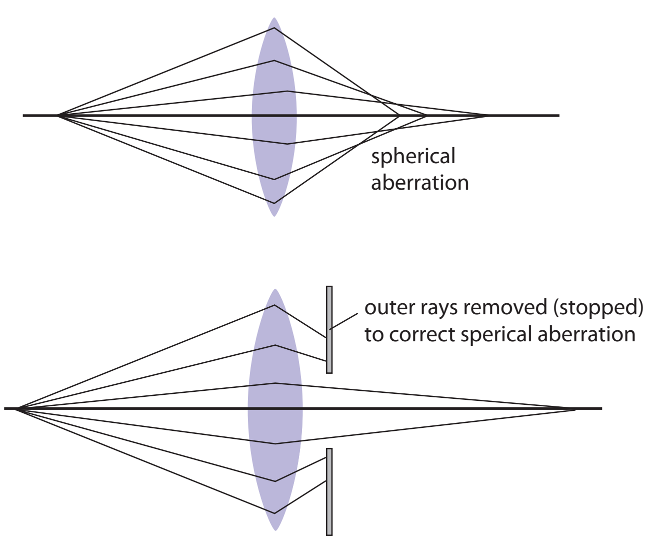
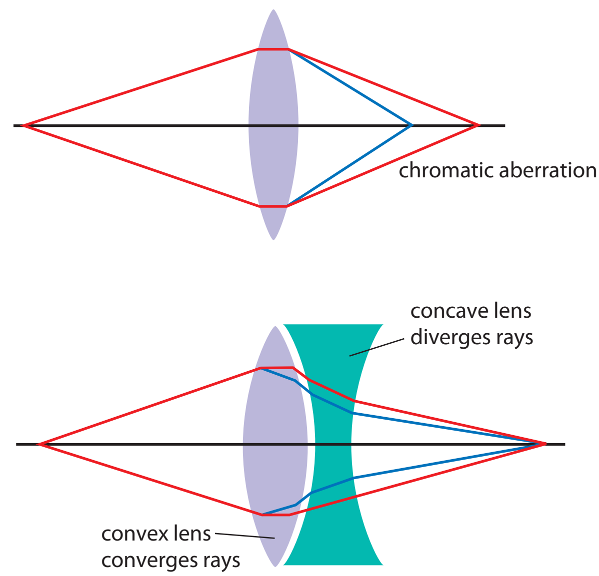


Figure 12.57 Spherical aberration in a convex lens.

Chromatic aberration

It has been mentioned before that different wavelengths of light are refracted by different amounts. If white light is focused with a convex lens the different colours are focused at different points. This also causes the image to be blurred. It can be corrected by making the lens out of two lenses of different refractive index stuck together. This is called an *achromatic doublet*. The blue light is most converged by the convex lens and most diverged by the concave one. These two effects cancel each other out.

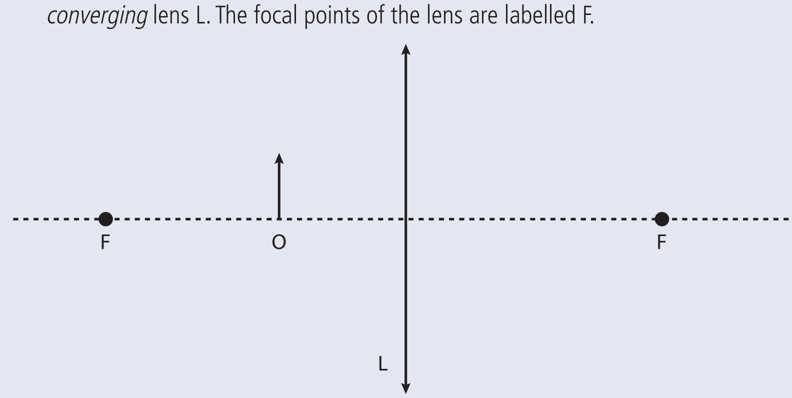
Figure 12.58 Chromatic aberration



Practice questions

1 This question is about converging lenses.

(a) The diagram shows a small object O represented by an arrow placed in front of a *converging* lens L. The focal points of the lens are labelled F.



(i) Define the *focal point* of a converging lens. (2)

(ii) On the diagram above, draw rays to locate the position of the image of the object formed by the lens. (3)

(iii) Explain whether the image is real or virtual. (1)

(b) A convex lens of focal length 6.25 cm is used to view an ant of length 0.80 cm that is crawling on a table. The lens is held 5.0 cm above the table.

(i) Calculate the distance of the image from the lens. (2)

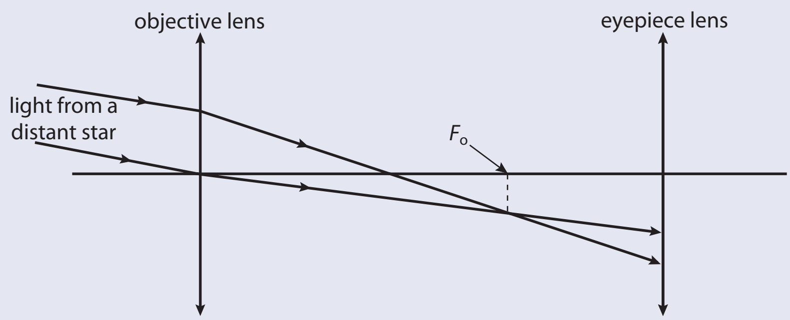
(ii) Calculate the length of the image of the ant. (2)

(Total 10 marks)

2 This question is about an astronomical telescope.

(a) Define the focal point of a convex (converging) lens. (2)

The diagram below shows two rays of light from a distant star incident on the objective lens of an astronomical telescope. The paths of the rays are also shown after they pass through the objective lens and are incident on the eyepiece lens of the telescope.



The principal focus of the objective lens is F_0

(b) On the diagram above, mark

(i) the position of principal focus of the eyepiece lens (label this F_E). (1)

(ii) the position of the image of the star formed by the objective lens (label this I). (1)

(c) State where the final image is formed when the telescope is in normal adjustment. (1)

(d) Complete the diagram above to show the direction in which the final image of the star is formed for the telescope in normal adjustment. (2)

The eye ring of an astronomical telescope is a device that is placed outside the eyepiece lens of the telescope at the position where the image of the objective lens is formed by the eyepiece lens. The diameter of the eye ring is the same as the diameter of the image of the objective lens. This ensures that all the light passing through the telescope passes through the eye ring.

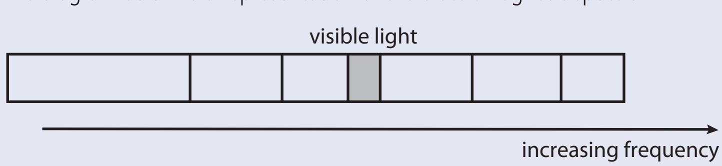
(e) A particular astronomical telescope has an objective lens of focal length 98.0 cm and an eyepiece lens of focal length 2.00 cm (i.e. $f_0 = 98.0$ cm, $f_e = 20.0$ cm). Determine the position of the eye ring. (4)

(Total 11 marks)

3 This question is about light and the electromagnetic spectrum.

(a) Outline the electromagnetic nature of light. (2)

(b) The diagram below is a representation of the electromagnetic spectrum.



In the diagram the region of visible light has been indicated. Indicate on the diagram above the approximate position occupied by

(i) infrared waves (label this I). (1)

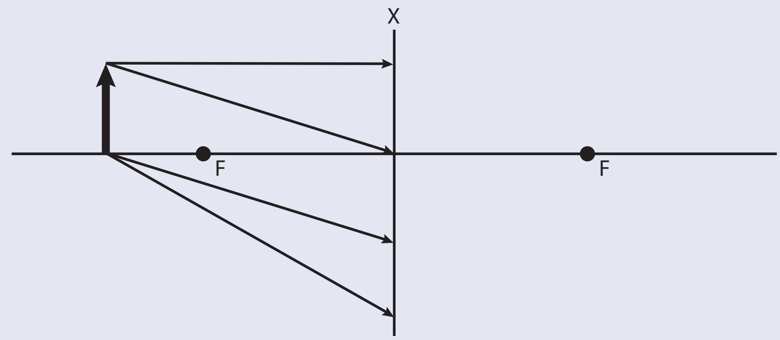
(ii) microwaves (label this M). (1)

(iii) gamma rays (label this G). (1)

(Total 5 marks)

4 This question is about a concave (diverging) lens.

The diagram below shows four rays of light from an object O that are incident on a thin concave (diverging) lens. The *focal points* of the lens are shown labelled F. The lens is represented by the straight line XY.



(a) Define the term *focal point*. (2)

(b) On the diagram,

(i) complete the paths of the four rays in order to locate the position of the image formed by the lens (4)

(ii) show where the eye must be placed in order to view the image. (1)

(c) State and explain whether the image is real or virtual. (2)

(d) The focal length of the lens is 50.0 cm. Determine the linear magnification of an object placed 75.0 cm from the lens. (3)

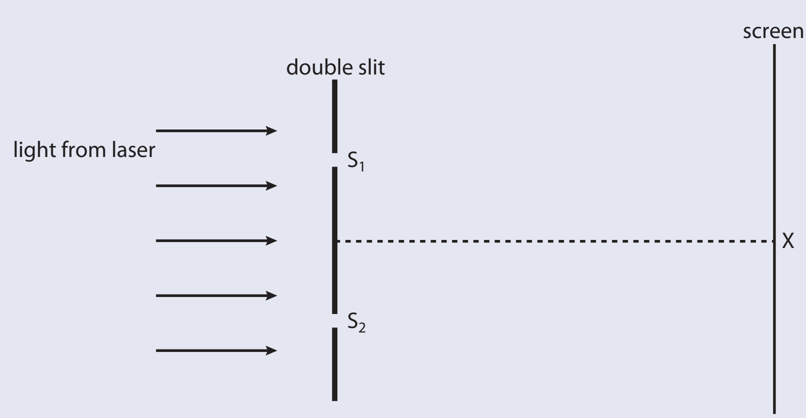
(e) Half of the lens is now covered such that only rays on one side of the principal axis are incident on the lens. Describe the effects, if any, that this will have on the linear magnification and the appearance of the image. (2)

(Total 14 marks)

5 This question is about the wave properties of light.

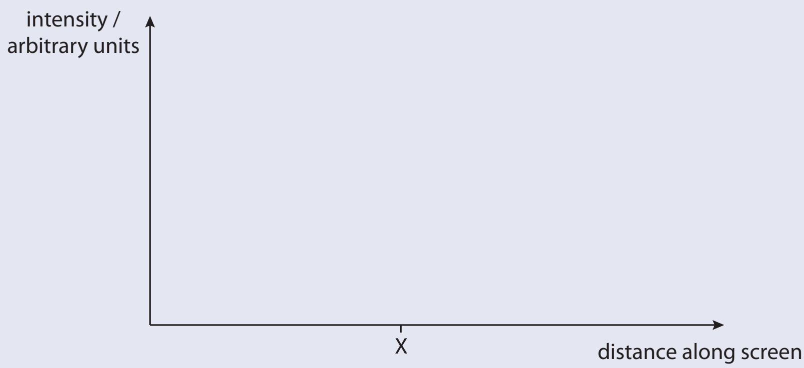
The diagram below (not to scale) is an arrangement for observing the interference pattern produced on a screen when the light from two narrow slits S_1 and S_2 overlaps. A beam of light from a laser is incident on the slits and after passing through the slits, the light is incident on a screen. The separation between the slits is large compared to the width of the slits and the distance between the slits and the screen is large compared to the slit separation.

The point X on the screen is equidistant from S_1 and S_2.



(a) Explain why an interference pattern will not be observed on the screen if the laser is replaced with a tungsten filament lamp. (2)

(b) On the axes below, draw a sketch-graph to show how the intensity of the observed interference pattern varies with distance along the screen. (2)

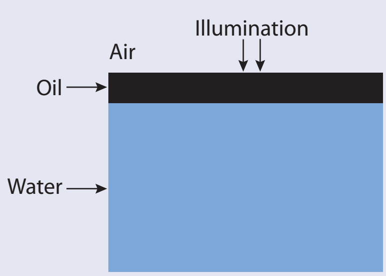


(c) The wavelength of the light from the laser is 633 nm and the angular separation of the bright fringes on the screen is 4.00×10^{-4} rad. Calculate the distance between S_1 and S_2. (3)

6 This question is about the formation of coloured fringes when white light is reflected from thin films.

(a) Name the wave phenomenon that is responsible for the formation of regions of different colour when white light is reflected from a thin film of oil floating on water. (1)

(b) A film of oil of refractive index 1.45 floats on a layer of water of refractive index 1.33 and is illuminated by white light at normal incidence.



When viewed at near normal incidence a particular region of the film looks red, with an average wavelength of about 650 nm. An equation relating this dominant average wavelength λ, to the minimum film thickness of the region t, is $\lambda = 4nt$.

(i) State what property n measures and explain why it enters into the equation. (2)

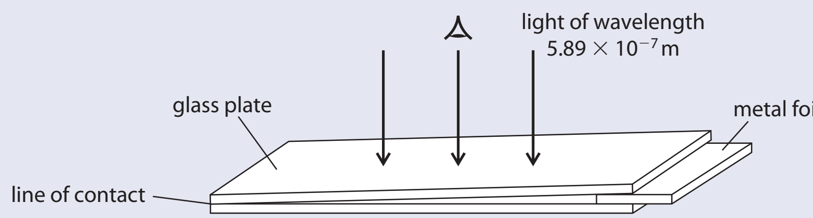
(ii) Calculate the minimum film thickness. (1)

(iii) Describe the change to the conditions for reflection that would result if the oil film was spread over a flat sheet of glass of refractive index 1.76, rather than floating on water. (2)

(Total 6 marks)

7 This question is about thin film interference.

Two flat glass plates are in contact along one edge and are separated by a piece of thin metal foil placed parallel to the edge, as shown below.



Air is trapped between the two plates. The gap between the two plates is viewed normally using reflected light of wavelength 5.89×10^{-7} m.

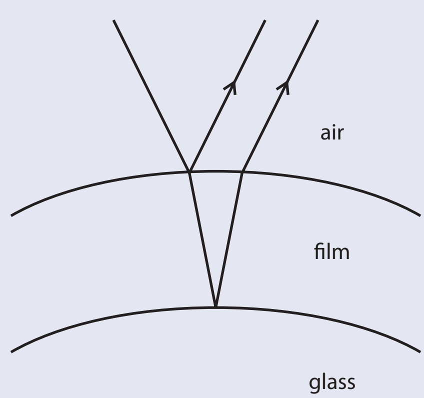
A series of straight fringes, parallel to the line of contact of the plates is seen.

(a) State what can be deduced from the fact that the fringes are straight and parallel. (1)

(b) Explain why a dark fringe is observed along the line of contact of the glass plates. (3)

(c) The distance between the line of contact of the plates and the edge of the metal foil is 9.0 cm. The dark fringes are each separated by a distance of 1.4 mm. Calculate the thickness of the metal foil. (3)

The lenses used in astronomical telescopes are frequently "bloomed". This means that a thin film is deposited on the lens in order to reduce the intensity of unwanted light reflected by the lens. Destructive interference occurs between the light reflected from the upper and the lower surfaces of the film. The reflections at both surfaces for one incident ray are shown in the diagram.



(d) **(i)** State why complete destructive interference of all the reflected light does not occur. (1)

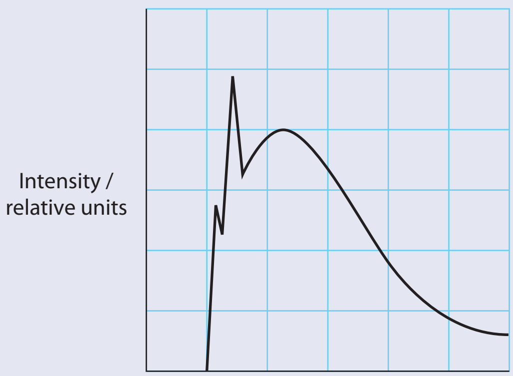
(ii) With reference to your answer in **(i)**, suggest why the film appears to be coloured. (2)

(Total 10 marks)

8 This question is about X-rays.

Electrons are accelerated through a potential difference of 25 kV and strike a target made from molybdenum.

The diagram below shows a partly labelled sketch graph of the resulting X-ray spectrum.



(a) On the diagram label

(i) the horizontal axis. (1)

(ii) a point P corresponding to the "braking radiation". (1)

(iii) the characteristic spectrum of the target. (1)

(b) Calculate the maximum frequency of the radiation produced. (2)

(Total 5 marks)

13 Relativity (Option H)

13.1 Special relativity

Assessment statements

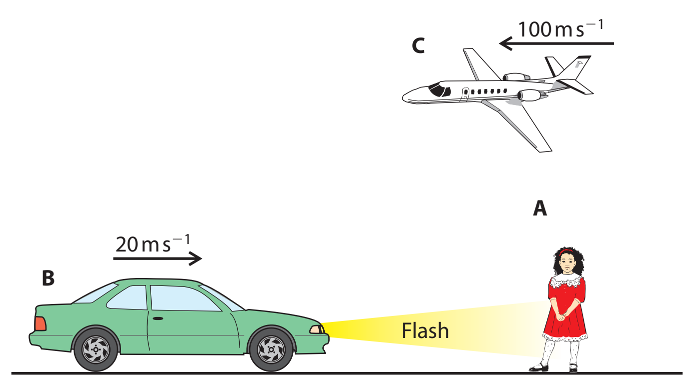
H.1.1 Describe what is meant by a frame of reference.
H.1.2 Describe what is meant by a Galilean transformation.
H.1.3 Solve problems involving relative velocities using the Galilean transformation equations.
H.2.1 Describe what is meant by an inertial frame of reference.

During this course we have sometimes used the term *relative*; for example, when quoting a velocity, it is very important to say what the velocity is measured relative to.

Relative velocity

Consider the example shown in Figure 13.1. **A**, **B** and **C** measure each other's velocity but they do not agree.

Figure 13.1 **A**, **B** and **C** measure each other's velocity.



Measured by A:

Velocity of car $= 20\,\text{m s}^{-1}$

Velocity of plane $= -100\,\text{m s}^{-1}$

Measured by B:

Velocity of girl $= -20\,\text{m s}^{-1}$

Velocity of plane $= -120\,\text{m s}^{-1}$

They don't agree because velocity is relative.

When considering an example like this there are some useful terms worth defining.

Event

An event is some change that takes place at a point in space at a particular time. If **B** were to flash his headlights, this would be an event.

Observer

An observer is someone who measures some physical quantity related to an event. In this case **A** measures the time and position when **B** flashed his lights, so **A** is an observer.

Frames of reference

A frame of reference is a coordinate system covered in clocks that an observer uses to measure the time and position of an event. It is covered in clocks so we can measure the time that an event took place *where* it took place. If we used our own clock we would always measure a time that was a little late, since it takes time for light to get from the event to us. An observer can only make measurements in their own frame of reference.

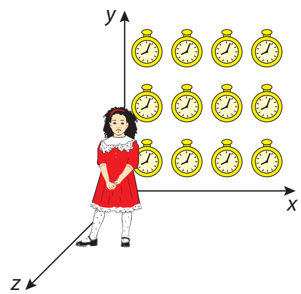


Figure 13.2 **A** in her frame of reference with some of her clocks. We won't always draw the clocks but remember they are there.

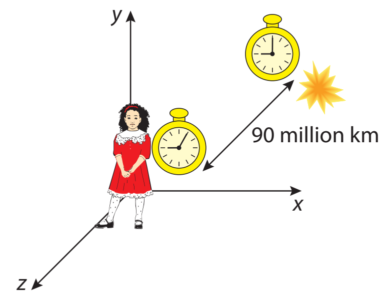


Figure 13.3 It takes 5 minutes for the light to reach **A** from this event. If she had used her own clock she would have got the wrong time.

In the example of Figure 13.1 we have 3 observers with 3 different frames of reference. As we can see in Figure 13.4, the frames of reference are moving relative to each other with constant velocity.

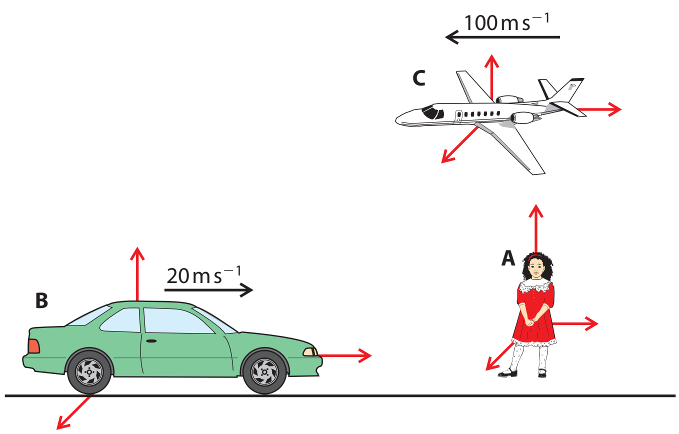


Figure 13.4 Three observers three frames.

When we look at Figure 13.4 we can see that the car and plane are moving, but the girl is standing still. This is because we often measure velocity relative to the Earth and she is stationary relative to the Earth. But, according to **B**, **A** is moving with a velocity of $20\,\text{m}\,\text{s}^{-1}$ to the left, so who is moving – **A** or **B**? Is there an experiment that we can do to prove which one of them is moving and which one is stationary?

Inertial frame of reference
An inertial frame of reference is a frame of reference within which Newton's laws of motion apply.

The presence of the Earth can be confusing, since we always think of the person standing as stationary. Also the Earth's gravitational field and the fact that it is spinning complicate matters. For that reason we will now move our observers into space, as in Figure 13.5, and ask the question again, is there any experiment that **A** or **B** could do to prove who is moving and who is stationary?

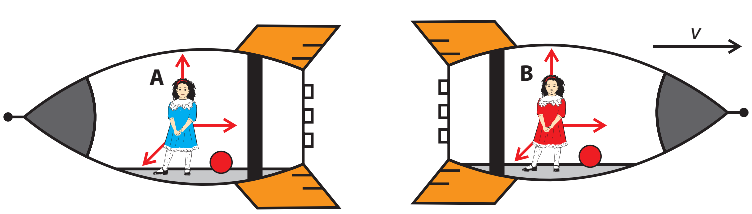


Figure 13.5 Two observers; **B** is travelling at velocity v relative to **A**.

Let's try a simple experiment. **A** and **B** take a ball (red) and place it on the floor. If they apply a force to the ball, it will accelerate and if they don't, it will remain at rest. Newton's laws of motion apply in each frame of reference. There is in fact no experiment that **A** and **B** can do to show who is moving – they are moving relative to each other but there is no absolute movement. We call these 'inertial frames of reference'. Let us compare this to the situation in Figure 13.6 where **B**'s rocket is accelerating.

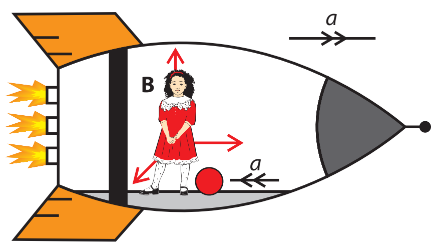


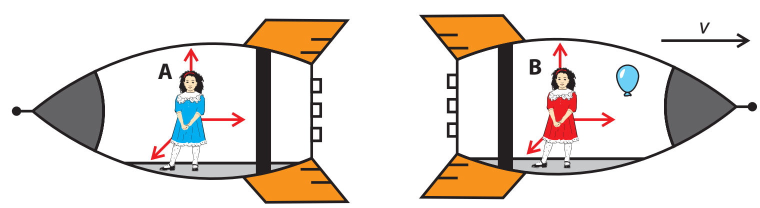
Figure 13.6 Rocket **B** fires the engines and accelerates.

If observer **B** now places a ball on the ground it will start to roll towards her even though there is no force acting on the ball, Newton's laws do not apply! Watching from the outside we can see what is happening, there is a force acting on the rocket pushing it past the ball and the ball is stationary. Newton's laws have not been broken, but inside the rocket it appears that they have. Accelerating frames of reference like this are called 'non-inertial frames of reference'.

Coordinate transformations

If observer **A** and **B** in Figure 13.7 both measure the position of the blue balloon floating weightlessly in the spaceship of **B** then they will get different values. So if **B** were to tell **A** where the ball was, it wouldn't make sense unless **A** knew how to transform **B**'s measurement into her own frame of reference. The classical way of doing this is called a 'Galilean transformation'.

Figure 13.7 A blue ball is at rest in **B**'s rocket.



Galilean transformations

At some time, as **A** and **B** are flying apart, the balloon bursts. **A** and **B** measure the time and position of this event. To do this they will use clocks and tape measures in their own frames of reference but to transform the measurements they need to have some reference point. So let's assume that at the time when they started their clocks, the two frames were at the same place. This isn't really possible with the rockets but we can imagine it.

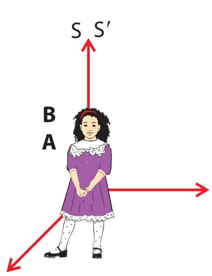
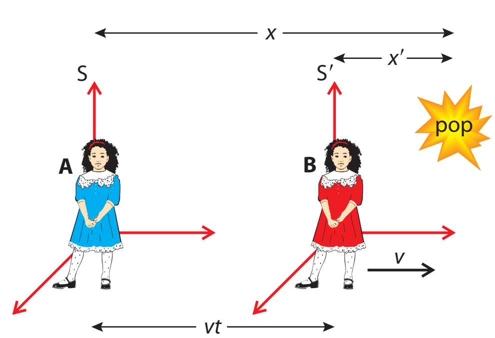


Figure 13.8 At time $t = 0$, **A** and **B** are coincident, as can be seen from the magenta dress.

If we call **A**'s frame of reference S and **B**'s S′ (S and S dash), we can then distinguish between **A** and **B**'s measurement by using a dash.

Figure 13.9 shows the moment that the balloon bursts.

Figure 13.9 The balloon pops when **A** and **B** have moved apart.



● **Hint:** The transforms work in both directions. Sometimes you might see examples where the event is at rest in **A**'s frame of reference rather than **B**'s.

A and **B** record the coordinates and time for this event and get the following results:

A (S)	**B** (S′)	Transformation
x	x'	$x = x' + vt$
y	y'	$y = y'$
z	z'	$z = z'$
t	t'	$t = t'$

We can see that only the coordinate in the direction of motion (x) is changed.

Galilean transforms for velocity

Galilean transformations can also be used to transform velocities. Consider a small bird flying in the x direction in **B**'s spacecraft as shown in Figure 13.10.

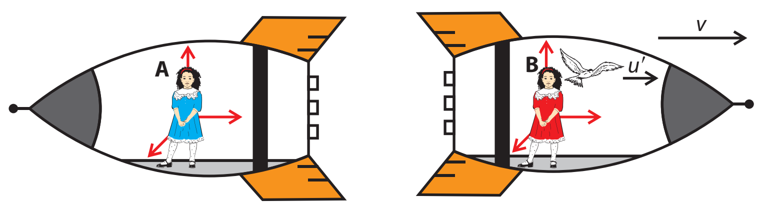


Figure 13.10 A bird flies with velocity u' as measured by **B**.

A (S)	**B** (S′)	Transformation
u	u'	$u = u' + v$

Hint: This is the same as subtracting the velocity of **A**'s frame relative to **B**.
$u' - -v = u' + v$

In this case, the velocity of the bird measured by **A** is found by adding the relative velocity of the frames of reference to the bird's velocity.

Exercises

1 A train travels through a station at a constant velocity of 8 m s^{-1}. One observer sits on the train and another sits on the platform. As they pass each other, they start their stopwatches and take measurements of a passenger on the train who is walking in the same direction as the train. Before starting to answer the following questions, make sure you understand what is happening; drawing a diagram will help.

(a) The train observer measures the velocity of the passenger to be 0.5 m s^{-1}. What is the velocity to the platform observer?

(b) After 20 s how far has the walking passenger moved according to the observer on the train?

(c) After 20 s how far has the walking passenger moved according to the observer on the platform?

The velocity of light

Assessment statements

H.5.3 Outline the Michelson–Morley experiment.
H.5.4 Discuss the result of the Michelson–Morley experiment and its implication.
H.5.5 Outline an experiment that indicates that the speed of light in vacuum is independent of its source.
H.2.2 State the two postulates of the special theory of relativity.
H.2.3 Discuss the concept of simultaneity.

Whenever we quote a velocity, we should say in which frame of reference we have measured that velocity. Some velocities have standard reference frames, so we don't need to do this; for example, the velocity of a car is relative to the road, the velocity of sound is relative to the air, and the velocity of a bullet is relative to the gun.

So, what is the velocity of light measured relative to?

How do you measure the velocity of light?

Light travels very fast (about 300 million $m\,s^{-1}$) so to measure its velocity you need to either use long distances or have a very fast clock. Hundreds of years ago clocks weren't fast enough, so methods were either based on astronomical events, where the distances travelled by light were large, or they made use of rotating shutters or mirrors to send pulses at known intervals around a fixed path. Today clocks are fast enough to measure time taken for light to travel a few centimetres, so there is no difficulty measuring its speed. Alternatively, since light is a wave, the velocity can be found by measuring its wavelength and frequency and using the formula

$$v = f\lambda.$$

Moving sources

If the velocity of light were relative to the source, then light from a source moving relative to an observer would have a different velocity to one that was stationary. Because the speed of light is very high, the change in velocity due to the movement of everyday objects like cars would be immeasurable, so very fast sources need to be used. Initial experiments were carried out using light from stars that were known to be moving relative to Earth. Although no difference was found, the results were not conclusive. In 1964 physicists working at the particle accelerator at CERN in Geneva measured the velocity of gamma ray photons emitted from neutral pions travelling at almost the speed of light. The results showed that even though the source was moving at $0.999c$ the light emitted still only travelled at c.

This result in itself is not particularly strange since sound would be the same; its velocity also does not increase if the source moves.

Moving medium

If the speed of a boat is $5\,m\,s^{-1}$ then it means that it will go through the water at $5\,m\,s^{-1}$, so if you stand on a river bank and measure the velocity of the boat going past in still water its velocity will be $5\,m\,s^{-1}$. If the water is moving at a velocity of $3\,ms^{-1}$ then the boat will have a velocity $8\,m\,s^{-1}$ if it moves in the direction of the water and $2\,m\,s^{-1}$ if it moves in the opposite direction.

Is the velocity of light affected by movement of its medium?

Since light travels throughout the universe, whatever medium it is that light travels through must fill the universe and unless we are at the centre of the universe we must be travelling through it too. So if the velocity of light is affected by movement of the medium, we should be able to detect a difference when light travels with the medium and against the medium. In 1881 Michelson and Morley devised an ingenious experiment to measure the effect of the medium on the velocity of light. To help understand this experiment we will first consider the analogous situation of a boat on a river.

The Michelson-Morley experiment with boats

This experiment is to verify that a boat's velocity is dependent on the velocity of the water. Consider two identical boats travelling to the red markers and back in still water, as shown in Figure 13.11

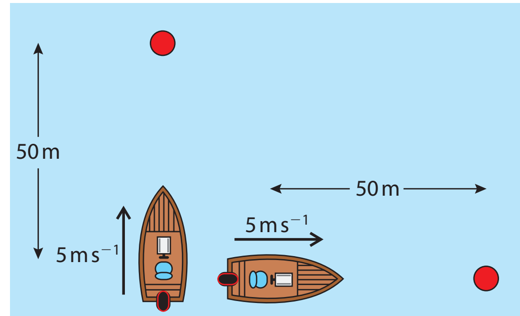


Figure 13.11 Boats make the journey in still water.

The time taken for each trip $= \dfrac{\text{distance}}{\text{speed}} = \dfrac{2 \times 50}{5} = 20$ s.

If the boats leave at the same time, they will return together.

If we now do the same trips in a river flowing at 3 m s^{-1} the boat travelling across the river will have to head up river with a velocity equal and opposite to that of the river, otherwise it will be swept downstream. The components of the boat's velocity are shown in Figure 13.12.

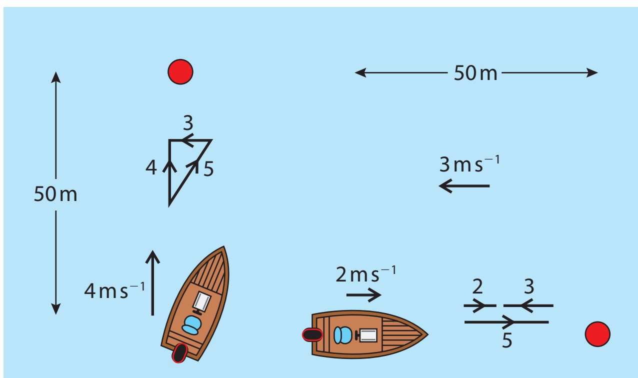


Figure 13.12 When the river flows the boat must head upstream so its velocity across the river is only 4 m s^{-1}.

Note

3, 4 and 5 m s^{-1} have been chosen here so that the triangle is an easy one. The vector triangle is showing the addition of 5 m s^{-1} and 3 m s^{-1} which add up to give the resultant velocity of 4 m s^{-1} across the river.

The boat is shown pointing up river but it will actually travel straight across.

The boat heading across the river will travel at 4 m s^{-1} there and back so will take $\dfrac{100}{4} = 25$ s.

The boat going upstream will travel at $5 - 3 = 2$ m s^{-1} on the way up, taking 25 s, and $5 + 3 = 8$ m s^{-1} on the way back, taking 6.25 s, a total of 31.25 s. The boat travelling across will therefore get back first.

If the boat's velocity were not changed by the movement of the water, boats would always arrive back together.

The Michelson-Morley experiment with light

The experiment with light uses a system of mirrors and a glass plate as shown in Figure 13.13.

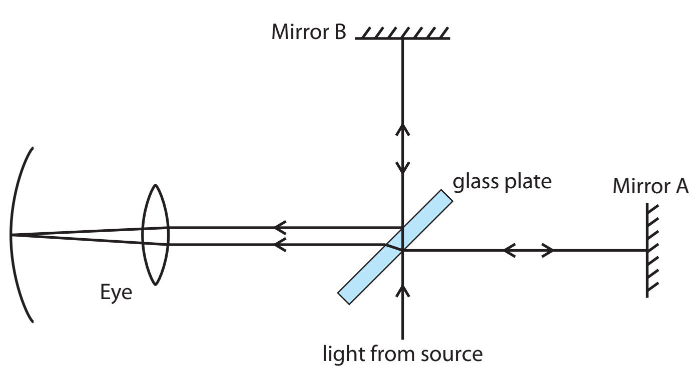


Figure 13.13 The Michelson-Morley experiment with no movement of medium.

When light hits the glass plate, some is reflected to **A** and some is refracted to **B**. Both rays are reflected back and will interfere when they meet in the eye.

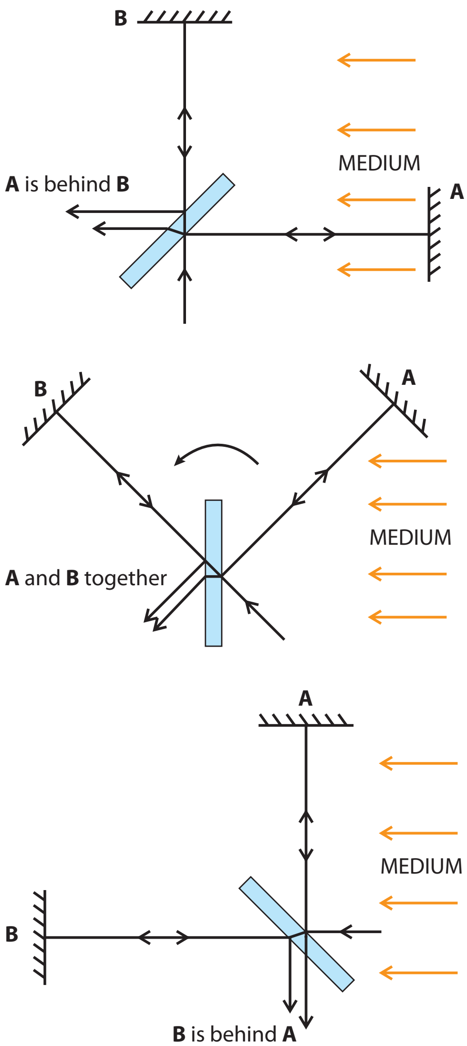


Figure 13.14 The result expected if the Michelson-Morley apparatus is rotated. This is **not** what happened.

If both paths are equal then there is no path difference so constructive interference takes place. If the medium is moving through the apparatus from right to left then the light travelling to **A** will take longer, resulting in a phase change when the light rays meet. Let's say the path difference $= \frac{\lambda}{2}$, then the waves will cancel. If the apparatus is rotated by 45°, the paths will be equal – constructive interference. Then after a further 45° they will cancel again.

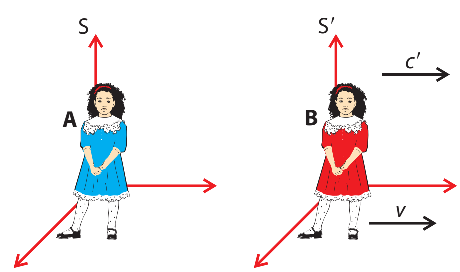
This is a simplification of the actual experiment but the principle was the same.

The result was that there was no detectable difference as the apparatus was rotated, showing that the velocity of light is not dependent on the movement of any medium, or that there is no medium.

The velocity of light is constant

The conclusion that we reach from the previous experiments is that the velocity of light in a vacuum is always the same as measured by all inertial observers, so no matter how fast I move towards or away from a source of light, the light will always approach me at the same speed. This causes problems if we try to apply the Galilean transformations to a photon of light.

Figure 13.15



If **B** measures the velocity of light to be c′ then according to the Galilean transformation, the velocity measured by **A** will be $c' + v$. This cannot be true, since the velocity of light must be the same to both observers – the transforms don't work. Einstein's theory of special relativity addresses this problem, resulting in some very interesting and far-reaching consequences.

Einstein's postulates

Einstein's postulates are the basis of the theory of relativity; if they are true, then all that follows must be true as well, even if it seems strange.

First postulate

The laws of physics are the same in all inertial frames of reference.

Second postulate

The speed of light in a vacuum is the same as measured by all inertial observers.

For a simulation of this experiment, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 13.1.

Historical context
The Michelson-Morley and pion experiments were not the basis of Einstein's postulates. However, they do support the theory. Einstein's original reason for thinking that the speed of light was constant was based on electromagnetic theory.

Simultaneity

The constancy of the speed of light has some interesting consequences when considering two events that happen at the same time (simultaneous events). Consider the train in Figure 13.16 travelling along a track at a relativistic speed. As the train passes, rabbit **B** sitting in a field, sees the train getting struck by lightning on both ends at the same time. Will rabbit **A** travelling in the train see the same thing?

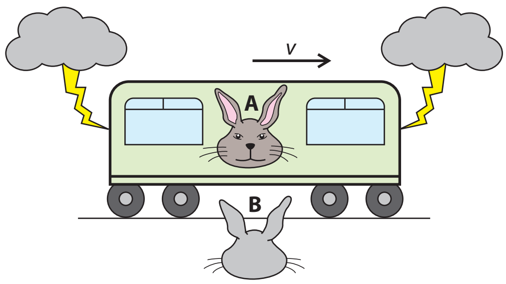


Figure 13.16 Two rabbits watch as a train gets struck by lightning. (They have to be rabbits, as rabbits have eyes on the side of their heads so can see both ends at the same time.)

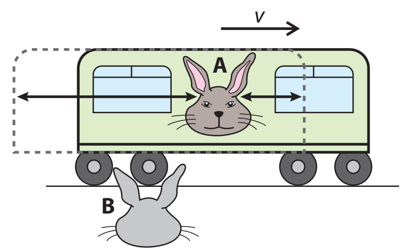


Figure 13.17 As the train progresses, the light from the front has a shorter path than the light from the back.

For rabbit **A** to see the lightning strikes, light from each strike must travel along the train towards the rabbit. As the light travels along the train, the train moves forward – this means that the light from the front will reach the rabbit before the light from the back, so rabbit **A** will see the front getting hit before the back. This is a consequence of the constancy of light because if light didn't have a constant speed, the light from the front would travel down the train more slowly. Its velocity would be $c - v$, and the light from the back would travel faster at $c + v$, resulting in both flashes arriving at the same time.

This problem can also be reversed; if the strikes are simultaneous to **A** then to **B**, the back was struck first.

Events at two different points in space that are simultaneous in one frame of reference are not simultaneous in all frames of reference.

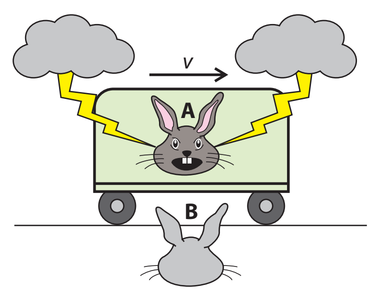


Figure 13.18 If the rabbit gets struck by two lightning bolts, then both will agree that the events are simultaneous since the events are at the same place.

13.2 Time dilation

Assessment statements

H.3.1 Describe the concept of a light clock.
H.3.2 Define proper time interval.
H.3.3 Derive the time dilation formula.
H.3.4 Sketch and annotate a graph showing the variation with relative velocity of the Lorentz factor.
H.4.1 Describe how the concept of time dilation leads to the "twin paradox".
H.4.2 Discuss the Hafele–Keating experiment.
H.3.5 Solve problems involving time dilation.

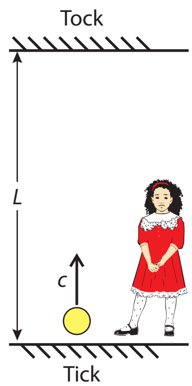


Figure 13.19 The light clock.

If the transforms are wrong, then maybe there is a simple way of fixing them. The simplest solution would be to multiply the transform by a constant, $x = k(x' + vt')$. To determine the value of k you would apply the condition that the speed of light had to be the same as measured in both frames of reference, $c = c'$. If you do this correctly, you find that this does have a solution, and that

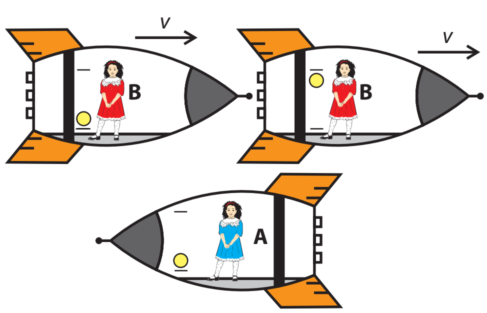
$$k = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

This is a rather lengthy derivation. A more visual approach is to consider an imaginary piece of equipment called a 'light clock'.

The light clock

A light clock is made from two parallel mirrors, as shown in Figure 13.19. All clocks are based on some physical event that always takes the same amount of time, for example the swinging of a pendulum. The light clock is based on the motion of a photon of light (the yellow ball) as it travels between the two mirrors. The distance travelled by the light between a 'tick' and a 'tock' is L so if light travels at speed c then the time $T_0 = \frac{L}{c}$.

Let us now put two identical light clocks in the two rockets considered previously.



Time dilation

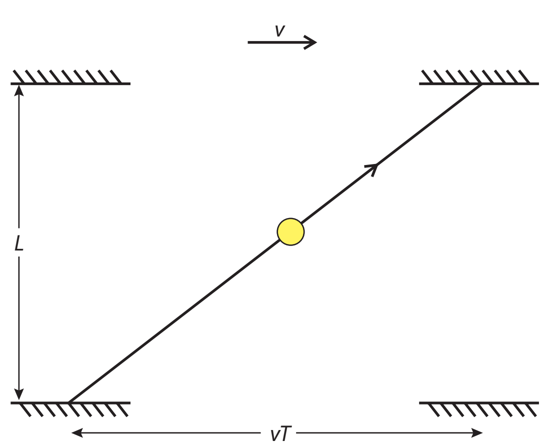
A will see her own light clock stationary with respect to herself but will see the light clock of **B** flying past. The distance travelled by the photon between a tick and a tock for her own clock is L. However, according to **A**, the photon in **B**'s clock not only travels across the clock but travels forwards at the same time. This means that **B**'s photon travels further. Since the speed of light is always the same, the time taken by **B**'s clock (as seen by **A**) will be longer.

Moving clocks tick slowly.

This is called *time dilation*.

To calculate how much more slowly the moving clock ticks, we need to know how long the path is. We can calculate this by looking at the triangle in Figure 13.20.

Figure 13.20 **B**'s clock as seen by **A**.



To view an animation of the light clock, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 13.2.

As **A** watches **B**'s clock, **A** will see it move forwards by a distance vT in the time between a tick and a tock. The length of the path can be found using Pythagoras. T is the time between Tick and Tock of **B**'s clock as measured by **A**. T_0 is the time of **A**'s own clock.

Distance travelled in time T, $cT = \sqrt{L^2 + v^2T^2}$

But L is the same as the length of **A**'s own clock so $L = cT_0$

Substituting gives $cT = \sqrt{c^2T_0^2 + v^2T^2}$

Squaring $c^2T^2 = c^2T_0^2 + v^2T^2$

Rearranging gives $T = \dfrac{T_0}{\sqrt{1 - \dfrac{v^2}{c^2}}}$

or if we define the constant γ as $\gamma = \dfrac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}}$

then $\boldsymbol{T = \gamma T_0}$

The constant γ is greater than 1 so $T > T_0$. The time taken for the photon to travel across the moving clock is longer than the time across the stationary one.

It is important to realise that this doesn't just apply to the light clock but everything. If that were not true, then you could do an experiment in one rocket that had a different result in the other, and that would be against Einstein's first postulate.

Lorentz transforms

The constant γ is called the Lorentz factor. The transforms for position and time that take into account the constant velocity of light are called the Lorentz transforms.

Proper time

Proper time is defined as a time measured by a clock at rest relative to the event. In this case T_0 is the proper time.

Variation of γ with v

Time dilation is only measurable when the relative velocity of the two inertial frames is very big. This is because the factor γ is almost 1 unless the v is close to c. This is illustrated by the graph in Figure 13.21.

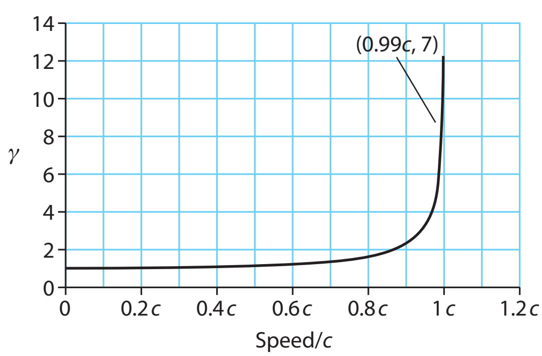


Figure 13.21 Graph of γ against speed.

● **Hint:** It is much more convenient to quote large velocities as a fraction of c rather than in m s^{-1} especially if you are going to calculate γ.

From this graph we can see that at low speeds, γ is approximately 1 (at 5×10^6 m s^{-1} for example, $\gamma = 1.000139$). As the speed approaches the speed of light, γ tends to infinity. A clock travelling past at the speed of light wouldn't tick, or to put it the other way round, if we travelled past the Earth at the speed of light, time on Earth would be stopped. The people on Earth, however, would not notice any difference.

Is there any point in saying what we would see if we travelled at the speed of light, since we can't?

Experimental verification

This seems all very strange, but there is supporting evidence. In 1971 Hafele and Keating placed very accurate caesium clocks on aeroplanes flying around the world westward and eastward, then compared the time to clocks that were on the Earth's surface. Classically all clocks should give the same time when they meet, no matter where they have been. According to relativity, since the clocks are moving relative to each other, they will show different times. Due to the Earth's rotation, all the clocks are moving relative to the centre of the Earth; so taking a non-rotating point at the centre of the Earth as our reference, the clock moving east will have the largest velocity, the one moving west, the smallest, and the one on the surface will be in between. This means that taking the centre of the Earth as a reference point, the eastward clock will tick slowest, the westward fastest and the one on the surface will be in between. So when the clocks are compared, the eastward clock will show a later time and the westward clock will show an earlier time than the stationary clock. And this is what Hafele and Keating found.

The twin paradox

According to special relativity, time is relative, so if observers **A** and **B** are moving relative to each other, observer **A** will see **B**'s watch ticking slowly and observer B will see **A**'s watch ticking slowly. There is nothing wrong with this, since they are in different frames of reference, but what if they come together? This is what the twin paradox is all about. Imagine twins **A** and **B** waving goodbye to each other as twin **B** begins a very fast journey into space.

According to A

● **Hint:** It isn't really realistic to talk about rockets and trains travelling at relativistic speeds. But they are often used in IB questions, so will be used here too.

A sees **B**'s clock tick more slowly than her own, so thinks that **B** is aging more slowly than she is. **A** therefore expects that when **B** returns, **B** will be the younger.

According to B

B sees **A**'s clock tick more slowly than her own so thinks **A** is aging more slowly. **B** therefore expects that **A** will be the younger when she returns.

The paradox is that they cannot both be the younger.

The answer lies in the fact that **B** has experienced two inertial frames of reference, one on the way out and another on the way back. It can be shown that an observer experiencing two different frames of reference will always age more slowly than one staying in the same frame. Hence **B** will be younger than **A**.

Exercises

2 A scientist constructs two boxes with lights on top that flash 5 times each minute. One box is left in the laboratory and the other is stuck to the side of a manned rocket that is sent into space. Some years later the rocket returns, flying past the laboratory at 0.9 c.

(a) According to the scientist

  (i) what is the time between flashes of his own box?

  (ii) what is the time between flashes of the box on the rocket?

(b) According to the astronaut

  (i) what is the time between flashes of his own box?

  (ii) what is the time between flashes of the box on the rocket?

(c) Why don't they agree?

3 A muon is a particle that decays into an electron with a half-life of about 2×10^{-6} s. Muons are formed when fast moving particles from the Sun (pions) interact with the atmosphere. The KE of the pions is transferred to the muons, giving them a velocity of about 0.99 c, relative to an observer on Earth.

(a) If the half-life of the muon is 2×10^{-6} s as measured in the frame of reference of the muon, what will it be to an observer on the Earth?

(b) 100 muons fly between two points on the Earth. If the time measured between the points by an observer on the Earth is 2.8×10^{-5} s, how many muons would you expect to arrive at the second point?

13.3 Length contraction

Assessment statements

H.5.1 Discuss muon decay as experimental evidence to support special relativity.
H.5.2 Solve problems involving the muon decay experiment.
H.3.6 Define proper length.
H.3.7 Describe the phenomenon of length contraction.
H.3.8 Solve problems involving length contraction.
H.4.3 Solve one-dimensional problems involving the relativistic addition of velocities.

To view a simulation of length contraction, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 13.3.

The example about muons at the end of the last section highlights another consequence of special relativity. It is quite straightforward to measure muons using two Geiger tubes placed one on top of the other. Muons pass through both tubes and travel so fast that they give signals from both tubes at almost the same time. Using this detector, it is possible to measure the number of muons at the top of a mountain compared to the bottom. However, time dilation leads to some inconsistencies; to illustrate this we will consider an example.

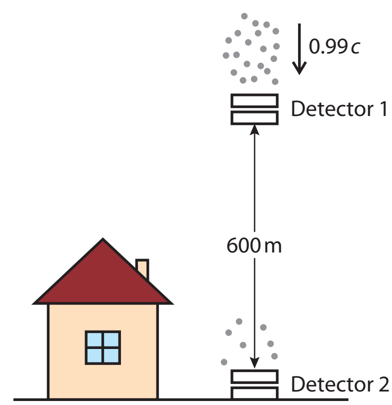


Figure 13.22

Muon decay

To make the calculation easier to follow, we will use the following approximate values:

Muon half-life measured at rest relative to muon = 2 μs

Speed of muon = 0.99 c so $\gamma = 7$

Distance between detectors = 600 m, as in Figure 13.22.

If a number of muons pass detector 1, what fraction arrive at detector 2?

In this example, the muons are used as a clock.

From the frame of reference of the muons:
The time taken for the muons to travel between the detectors is $\frac{600}{0.99\,c} = 2\ \mu\text{s}$.
This is one half-life, so half of the muons will arrive at the ground.

From the frame of reference of the ground:
The time taken for the muons to reach the ground is the same as before, 2 μs.

But to an observer on the ground, the muons' clock is ticking slowly, so, according to the Earth clocks, the half-life of muons is longer.

$$T = T_0\gamma = 2\ \mu\text{s} \times 7 = 14\ \mu\text{s}$$

So in the 2 μs taken for them to reach the Earth, very few will decay, leaving a lot more than half.

This result doesn't make sense; there can't be half and more than half left. The only way round this is if the distance travelled by the muons with their slow-ticking clocks was less than that measured on Earth. This is length contraction.

	Muon frame	Ground frame
Half-life	2 μs	14 μs
Distance	85.7 m	600 m
Time	0.28 μs	2 μs
Number of half-lives	0.14	0.14

Table 1 This table summarizes the results of the muon experiment from both frames of reference showing how they agree.

Before deriving the length contraction formula, we should sort out the muon problem:

To make the results consistent, the muons with their clocks (that to them are ticking 7 times faster than when viewed from Earth) must travel a distance that is only one-seventh the distance as measured by an observer on Earth.

So for the muons, the distance travelled is $\frac{600}{7} = 85.7$ m. If they travel at 0.99 c this will take 0.28 μs, which is much less than a half-life, so more than half survive, and the results agree.

If L_0 is the length measured by an observer at rest relative to the end points, then the length L measured by an observer moving relative to them is given by

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

Length contraction

To derive this formula, consider lines drawn on a road at X and Y as shown in Figure 13.23.

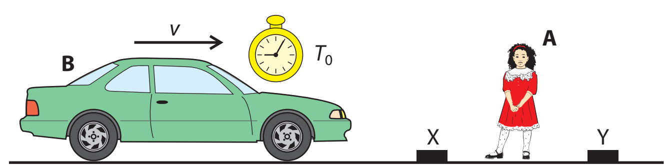


Figure 13.23 Observers **A** and **B** measure the distance between X and Y.

Observer **A**, who is at rest relative to the road, is going to measure the distance L_0 between the points, by timing how long it takes for the car driven by observer **B** to pass the points. The only clock is the clock belonging to **B**. This clock is moving relative to **A**, so the time T for the car to drive from X to Y is a relative time.

The car's speed is v so the distance measured by **A**

$$L_0 = vT$$

Observer **B** who is moving relative to the points also measures the distance, L, using the same clock

$$L = vT_0$$

Now we know that $T = \gamma T_0$ so $L_0 = v\gamma T_0$

Dividing these equations gives $\frac{L_0}{L} = \gamma$

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

So moving objects are shorter – lengths contract.

So if the velocity of the car is 0.99 c ($\gamma = 7$) and the distance measured by **A** is 2.1×10^9 m, the time measured by **A** (on **A**'s clock) will 7s. But if **A** watches **B**'s clock, she will see that it has only progressed 1s in the time for the car to travel from X to Y. That is because according to **B**, he has only travelled one-seventh of the distance.

Length contraction only applies to lengths that are in the direction of motion. The girl's height, for example, will be the same to both observers.

Proper length

Proper length is the length measured at rest relative to the object. In the previous case, that's L_0.

Simultaneity again

Another interesting example is that of a man carrying a ladder past a gate at relativistic speed. Measured at rest, the ladder and gate are the same length, but as the man runs past the gate he thinks the gate is shorter than the ladder, so won't fit through. An observer by the gate thinks the ladder is shorter, so it will fit through. Does it fit or not? To answer this, you need to realise that the events of each end passing through the gate do not have to be simultaneous to both observers. The man by the gate will see both ends passing at the same time. The observer with the ladder will see the front go through first then the back, so it does go through for both observers.

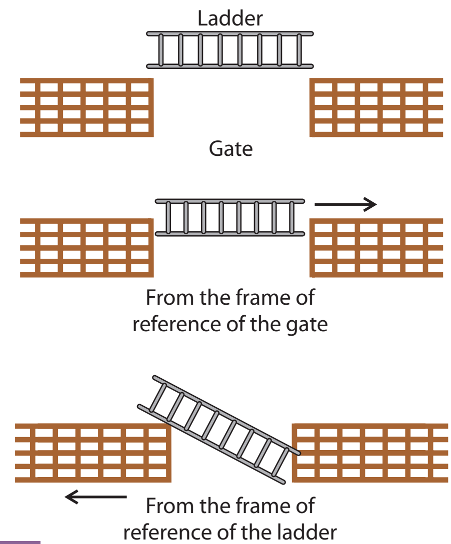


Figure 13.24 A ladder is pushed through a gate as it passes.

Exercises

4 Two relativistic knights as shown in Figure 13.25 charge past each other with their lances (spears) lowered, at a relative velocity of 0.5 *c*.

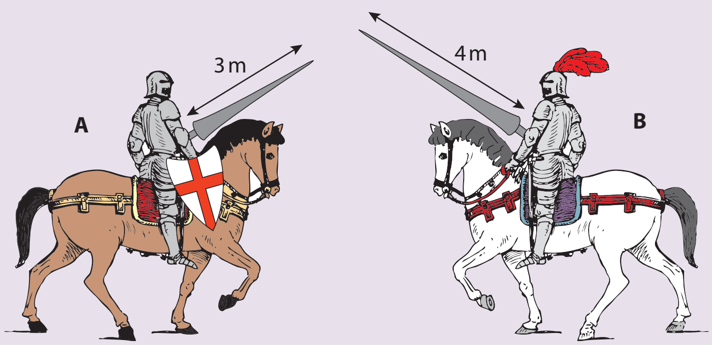


Figure 13.25 Relativistic knights in armour.

Each lance has been measured at rest, as written on the diagram.

(a) What is the length of **A**'s lance as measured by **B**?

(b) What is the length of **B**'s lance as measured by **A**?

(c) How fast would they have to go past each other for knight **A** to have a lance that was the same length as **B**'s?

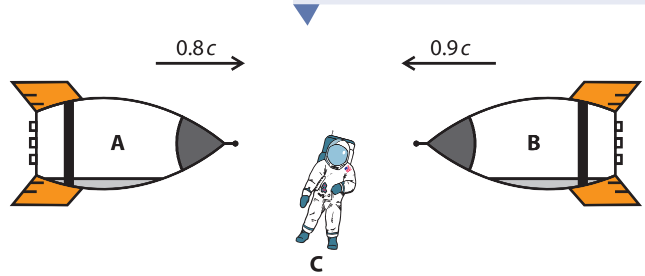
5 The distance between the Earth and the Sun is 1.5×10^{11} m. How far would this distance be as measured in the frame of reference of a particle travelling at 0.999 *c*?

6 How fast would an observer have to travel past an object so that it was half its proper length?

Velocity transformations

Using the Galilean transforms there is a problem when transforming velocities. For example, consider two rockets and a floating spaceman, as shown in Figure 13.26. The spaceman measures the velocity of rocket **A** to be 0.8 *c* and rocket **B** to be −0.9 *c*. If we use the Galilean transform to calculate the velocity of rocket **B** as measured by rocket **A** we get an answer of −1.7 *c*. This can't be right because **B** can't travel faster than the speed of light.

Figure 13.26 Two rockets with a closing speed of 1.7 *c* according to Galiliean transforms.



● **Hint:** It is important to realise that there are three possible frames of reference in this problem – the spaceman and the two spaceships. It's easy to forget when solving a problem about two spaceships, that there must be an outside observer observing the spaceships.

What we need is a Lorentz transform that will transform the velocity measured in one frame to another. First we should carefully define which are the frames of reference and which are the objects we are measuring.

Figure 13.27 Frames of reference defined.

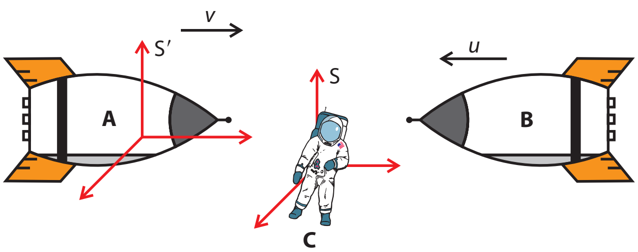


Figure 13.27 clarifies the frames of reference in this problem. The velocity of **B** measured in S is u, but we want to know its velocity measured in S′ which has a velocity v relative to S.

The derivation is beyond this course but if we apply the Lorentz transforms for position and time to this problem, we get the equation

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

If we substitute the values from the rocket example we get

$$u = -0.9\,c$$

$$v = 0.8\,c$$

$$u' = \frac{-0.9\,c - 0.8\,c}{1 - \frac{-0.9\,c \times 0.8\,c}{c^2}}$$

$$u' = \frac{-1.7\,c}{1.72}$$

$$\boldsymbol{u' = -0.988\,c}$$

● **Hint:** If **B** is a photon of light travelling at c, its velocity measured by **A**, will be

$$c' = \frac{-c - 0.8c}{1 - \frac{-0.8cc}{c^2}} = -c$$

So, the rockets do not approach each other faster than the speed of light. If the velocities are small then $\frac{uv}{c^2}$ is approximately zero, so the equation will be the same as the Galilean transform, $u' = u - v$.

Exercises

7 Two subatomic particles are collided head on in a particle accelerator. Each particle has a velocity of 0.9 c relative to the Earth. Calculate the velocity of one of the particles, as measured in the frame of reference of the other.

8 An observer on Earth sees a meteorite travelling at 0.5 c on a head-on collision course with a spaceship travelling at 0.6 c. What is the velocity of the meteorite as measured by the spaceship?

9 A relativistic fly flies at 0.7 c in the same direction as a car travelling at 0.8 c. According to the driver of the car, how quickly will the fly approach the car?

13.4 Mass and energy

Assessment statements

H.4.4 State the formula representing the equivalence of mass and energy.
H.4.5 Define *rest mass*.
H.4.6 Distinguish between the energy of a body at rest and its total energy when moving.
H.4.7 Explain why no object can ever attain the speed of light in a vacuum.
H.4.8 Determine the total energy of an accelerated particle.
H.6.1 Apply the relation $p = \gamma m_o u$ for the relativistic momentum of particles.
H.6.2 Apply the formula KE $= (\gamma - 1)\,m_o c^2$ for the kinetic energy of a particle.
H.6.3 Solve problems involving relativistic momentum and energy.

We have seen how the constancy of the speed of light changes length and time, and how, as a body gets faster, clocks tick slowly and objects contract. But how does this affect momentum and energy?

Relativistic mass

Imagine two Mexican boxing twins passing each other on two trains travelling at relativistic speeds as in Figure 13.28.

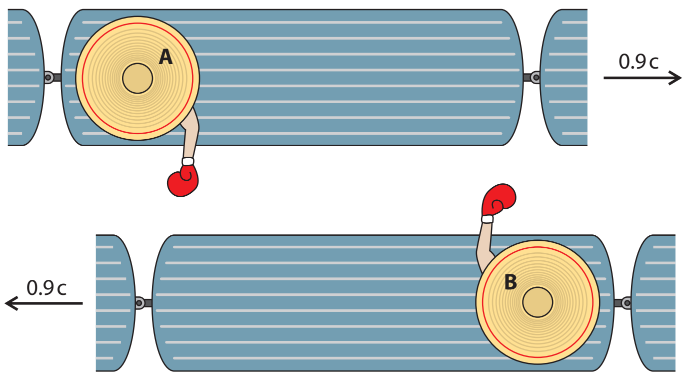


Figure 13.28 Two Mexican boxers passing each other on two trains as seen from above.

A Mexican and his hat.

As the trains approach, the boxers prepare to hit each other perpendicular to the direction of motion (if they hit each other in the direction of motion they would knock each other's heads off).

From A's frame of reference:

Due to time dilation, **A** will see **B**'s clock ticking slowly so everything that **B** does will be in slow motion. He laughs as he thinks how his punch will hurt much more than **B**'s

From B's frame of reference:

B sees **A**'s clock ticking slowly so thinks exactly the same thing.

When they pass each other, the twins are both surprised by the fact that the other twin's punch can hurt so much even though their glove was moving so slowly.

The only way for a slow-moving glove to have the same effect as a fast one is if it has a bigger mass. The mass of the moving glove is bigger than the slow one. It can be shown that the mass increases according to the following equation

$$m = \gamma m_0$$

where
m = the mass measured moving relative to the body, relativistic mass
m_0 = the mass measured at rest relative to the body, the **rest mass**
γ = the Lorentz factor.

Relativistic energy

We know that energy is transferred when work is done, and work is done when the point of application of a force moves in the direction of the force. If a constant unbalanced force is applied to a body, that body will accelerate ($F = ma$) gaining KE.

The mass – energy equivalence formula

$E = mc^2$

Travelling at the speed of light
We can now see why it is not possible for a body with mass m to travel at the speed of light. As the body goes faster and faster, it gets heavier and heavier. As we approach the speed of light, the energy needed to make it go faster becomes greater and greater. To reach the speed of light would require an infinite amount of energy.

The electronvolt
This is the amount of KE gained by an electron accelerated through a p.d. of 1 V. Masses can be expressed in $\text{MeV}c^{-2}$.

Mass of an electron $= 0.511\ \text{MeV}c^{-2}$.

Mass of a proton $= 938\ \text{MeV}c^{-2}$.

However, we have now discussed that as the body accelerates, its mass will increase from m_0 to m.

We have already come across the relationship between mass and energy in nuclear physics; we know that $E = mc^2$. So, if the increase in mass is Δm, then the amount of work done $\Delta E = \Delta mc^2$. The increase in mass is the change in mass from when it was at rest to its present velocity, which is $m - m_0$.

The amount of work done $= \text{KE} = (m - m_0)c^2 = mc^2 - m_0c^2$

where mc^2 = total energy

m_0c^2 = rest energy

work done = increase in KE

so KE = total E − rest E

This seems to make sense, since the energy a body has due to its movement equals the total energy of the body minus the energy it has when it's not moving.

But $m = \gamma m_0$

so $$\text{KE} = (\gamma m_0 - m_0)c^2 = (\gamma - 1)\, m_0c^2$$

So energy is in fact mass; when we do work we increase the mass of a body. However, at normal velocities this increase is not measurable.

Low velocity approximation

It is interesting to see what result the relativistic equation gives if the velocity of a body is much less than c.

$$\text{KE} = (\gamma m_0 - m_0)c^2 = \frac{m_0c^2}{\sqrt{1 - \frac{v^2}{c^2}}} - m_0c^2$$

Using the binomial expansion when x is small $(1 + x)^n = 1 + nx$ we can expand the term $\left(1 - \frac{v^2}{c^2}\right)^{-\frac{1}{2}} = 1 + \frac{v^2}{2c^2}$

So $\text{KE} = \left(1 + \frac{v^2}{2c^2}\right)m_0c^2 - m_0c^2 = \frac{1}{2}m_0v^2$

This is what we expected.

Units

Before doing any examples we will save a lot of time by thinking about the units we use. In real examples, trains and knights do not travel at relativistic speeds. To get something that really goes this fast, we need to accelerate charged particles using an electric field. When we do this, it is much more convenient to use MeV as the unit of energy rather than J. If we use the formula $E = mc^2$, we can calculate the energy equivalent of a mass. If the energy is in MeV then the mass can be given directly as $\text{MeV}c^{-2}$. For example, if we convert an electron into energy we get 0.5 MeV and the mass of the electron is therefore $0.5\ \text{MeV}c^{-2}$. If we also use fractions of c as the unit of velocity, calculation becomes quite straightforward.

Worked examples

1 Calculate the speed of an electron that is accelerated through a p.d. of 1 MV.

2 Calculate the p.d. required to accelerate an electron to a velocity of 0.8 c.

Solutions

1 Loss of electrical PE = gain in KE of electron

so KE = 1 MeV

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

where m_0c^2 is the rest energy of an electron = 0.5 MeV

So $1.0 = (\gamma - 1) \times 0.5$

Rearranging gives $\gamma = 3.0$

But $\gamma = \dfrac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}} = 3.0$

$v = 0.94\,c$

2 If $v = 0.8\,c$ then $\gamma = 1.67$

If the electron is travelling at $0.8\,c$ it will have a KE $= (\gamma - 1)m_0c^2$

where m_0c^2 = rest energy

= 0.5 MeV

KE $= (1.67 - 1) \times 0.5 = 0.34$ MeV

So the p.d. must have been 0.34 MV

Exercises

10 Calculate the speed of a proton accelerated through a p.d. of 1 MeV.

11 Calculate the total energy of an electron moving with a velocity of $0.9\,c$.

12 What p.d. would be required to give an electron a total energy of 2.5 MeV?

13 What p.d. is required to give a proton a velocity of $0.7\,c$?

Relativistic momentum

The momentum of a body is defined by the equation $p = mv$ but we have seen that the mass of a body is dependent on its velocity, $m = \gamma m_0$ so to take this into account we should write:

$$p = \gamma m_0 v$$

The momentum and energy of a body are very much connected; if you do work on a body, you increase its energy and also its momentum. For non-relativistic bodies:

$$\text{KE} = \tfrac{1}{2}mv^2 \text{ and } p = mv$$

So the relationship between momentum and energy is:

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

For relativistic bodies, the relationship is a bit more complicated:

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

where E = total energy

For a photon, $m_0 = 0$ so $E^2 = p^2c^2$

Rearranging gives $E = pc$

We know that the energy of a photon $= \dfrac{hc}{\lambda}$

so $\dfrac{hc}{\lambda} = pc$

Rearranging gives $p = \dfrac{h}{\lambda}$

So even though a photon has no mass, it still has momentum.

Momentum of a photon

Expanding the equation $p = \gamma m_0 v$

we get $p = \dfrac{m_0 v}{\sqrt{1 - \dfrac{v^2}{c^2}}}$

If the mass of a particle is zero then this can only have a solution if $v = c$. This will make the equation $p = 0/0$ which has a possible solution but it is indeterminate. This is consistent with photons having momentum even though they have no mass.

Units again

To keep things simple we can express momentum in $\text{MeV}c^{-1}$

So if we look at the equation

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

we have $$(\text{MeV})^2 = (\text{MeV}c^{-2})^2c^4 + (\text{MeV}c^{-1})^2\, c^2$$

This all balances.

Worked examples

1 What is the momentum of an electron with KE = 1 MeV?

2 What is the speed of an electron with a momentum of $2\,\text{MeV}c^{-1}$

Solutions

1 Total energy = rest energy + KE

$$= 0.5 + 1 = 1.5\,\text{MeV}$$

So $$E^2 = m_0{}^2c^4 + p^2c^2$$
$$1.5^2 = 0.5^2 + p^2c^2$$
$$p = 1.41\,\text{MeV}c^{-1}$$

2 First we can find the total energy from

$$E^2 = m_0{}^2c^4 + p^2c^2$$
$$E^2 = 0.5^2 + 2^2$$

So $E = 2.06\,\text{MeV}$

Total $E = \gamma m_0c^2$

so $$2.06 = \gamma \times 0.5$$
$$\gamma = 4.12$$
$$v = 0.97c$$

Exercises

14 Find the momentum of a particle of rest mass $100\,\text{MeV}c^{-2}$ travelling at $0.8c$.

15 A particle of rest mass 200 MeV is accelerated to a KE of 1 GeV. Calculate its
(a) momentum
(b) velocity.

16 A particle of rest mass 150 MeV is accelerated to a speed of $0.8\,c$. Calculate its
(a) KE
(b) total energy
(c) momentum.

17 A proton has a momentum of $150\,\text{MeV}c^{-1}$.
Calculate its
(a) total energy
(b) KE
(c) accelerating potential
(d) velocity.

13.5 General relativity

Assessment statements

- H.7.2 Describe and discuss Einstein's principle of equivalence.
- H.7.1 Explain the difference between the terms gravitational mass and inertial mass.
- H.7.3 Deduce that the principle of equivalence predicts bending of light rays in a gravitational field.
- H.7.4 Deduce that the principle of equivalence predicts that time slows down near a massive body.
- H.8.1 Outline an experiment for the bending of EM waves by a massive object.
- H.8.2 Describe gravitational lensing.
- H.8.3 Outline an experiment that provides evidence for gravitational red-shift.
- H.7.12 Describe the concept of gravitational red-shift.
- H.7.13 Solve problems involving frequency shifts.
- H.7.14 Solve problems using the gravitational time dilation formula.

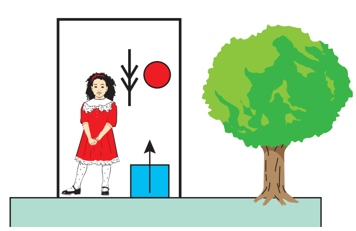
Special relativity only relates to inertial frames of reference. General relativity extends this to non-inertial frames, that is, a frame of reference within which an object will accelerate without being pushed. We have come across two examples of this, one was an accelerating rocket and the other is on the surface of the Earth. In both cases, a ball will accelerate downwards if released. The important thing to realise is that it is impossible to distinguish between these two examples; this is the starting point to general relativity.

The principle of equivalence

Einstein's principle of equivalence states that:

No observer can determine by experiment whether they are in an accelerating frame of reference or a gravitational field.

This may seem a bit odd, since it is very easy to tell whether you are going round a bend in a car, or falling off a cliff; so to make the point clearer we shall consider two observers in identical boxes, as in Figure 13.29. One box is sitting on the Earth, the other is accelerating with acceleration 9.81 m s^{-2} far out in space. All experiments must be done in the box and it's not allowed to look out of the window.



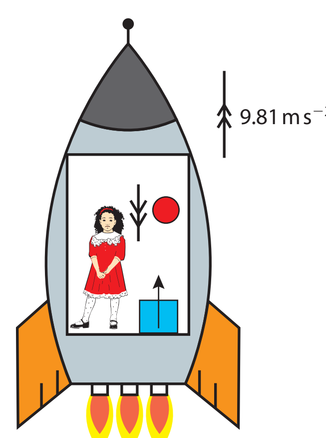


Figure 13.29 Two different but indistinguishable non-inertial frames of reference.

Whatever the two observers do they will get the same result:

The dropped red ball will fall with an acceleration *g*.

The blue box will sit on the floor experiencing a force, *N*, pushing it up. From the outside we can see they are quite different; for example the blue box on the Earth has another force acting on it, its weight, and these forces are balanced. The box in the rocket has only one force, which is unbalanced, so this box is accelerating. However, from inside the box you can't tell any difference.

Gravitational mass and inertial mass

In both of frames of reference in Figure 13.29, it is possible to measure the mass of one of the objects. On the Earth the force of gravity can be used to find mass by measuring the object's weight and calculate the mass from $F = mg$. This is **gravitational mass.**

In the spaceship, the force that must be applied to make an object accelerate with the spaceship can be measured, and the mass calculated using the formula $F = ma$. This is **inertial mass.**

These two quantities are the same.

Constant velocity
If the boxes travelled at constant velocity, the light would also leave each box at a different point but if you join these points you will get a straight line.

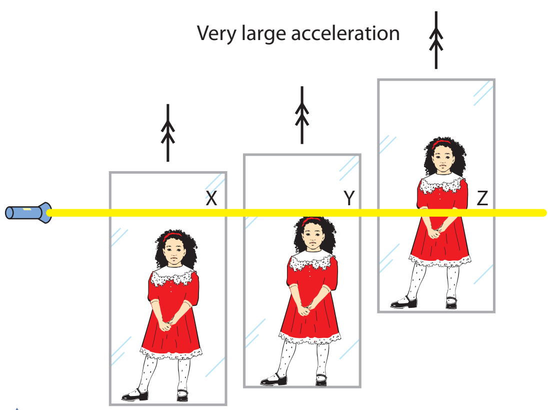


Figure 13.30 Three glass boxes photographed as the light leaves each box, The entry points would all be higher since the box has moved in the time for the light to travel across it.

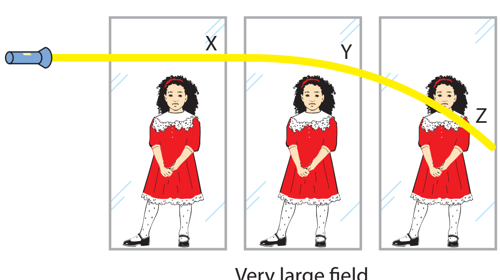


Figure 13.31 For the light to exit each box at the same place in a *G* field the light must have followed a curved path.

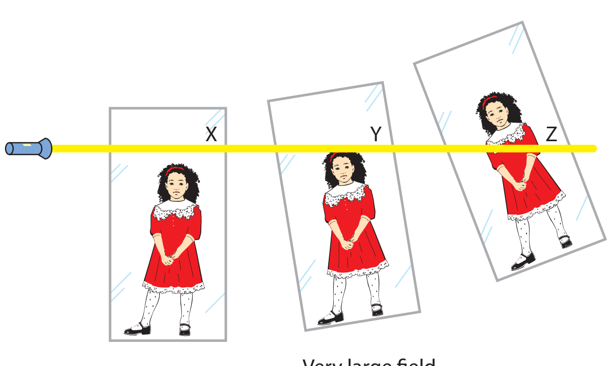


Figure 13.32 Instead of the light bending, maybe it's the space and time that it moves through which is bent.

Bending of light by gravity

The principle of equivalence has some far-reaching consequences when we consider what happens to light in an accelerating frame of reference, and realise that the same thing must happen in a gravitational field. If the girl in the rocket shone a beam of light across a box accelerating at only 9.8 m s^{-2} nothing much would happen; the beam would simply go straight across the room. However, if we increased the acceleration to astronomic levels we would get a different result. So that we can see what happens at different times we will consider three glass boxes starting together and accelerating at the same rate as shown in Figure 13.30. A beam of light is shone through all three boxes as they accelerate up.

As we can see from the diagram the boxes move up as the light passes through and because the boxes are accelerating, the last box has moved up more than the middle one. X, Y and Z are the positions where the light leaves each box. According to the principle of equivalence, the same thing should happen in a gravitational field. This time we would need to take the boxes to a very large field close to a very large mass but small radius $\left(g = \frac{GM}{r^2}\right)$. For the same thing to happen the light must exit each box at the same point as shown in Figure 13.31.

This shows that light is bent by the large mass but this doesn't fit into our accepted theory of gravity. Gravitational force is proportional to mass and since light has no mass, how can it be attracted to the Earth in this way? However, there could be another explanation as shown in Figure 13.32.

The solution illustrated in Figure 13.32 may seem a bit far-fetched and it isn't really correct, but it does give an idea of how Einstein's theory of general relativity tackles the problem by abandoning the idea of our coordinate system having straight lines, and considering what would happen if space–time was curved.

Large masses curve space–time.

Bending of light by the Sun

If light is bent by objects with large mass then light should be bent as it passes the Sun. This is rather difficult to test because the Sun gives out so much light itself that you can't see light coming from behind, except during an eclipse. During a total eclipse it is possible to see stars that you normally only see at night. However, some of those stars will be behind the Sun so won't be visible. During the eclipse of 1919 Arthur Eddington found that the positions of some stars that should have been close to the Sun were shifted outwards. This could be explained if the Sun bends the light from the stars as in Figure 13.33.

One of Arthur Eddington's original photographs from the solar eclipse in 1919.

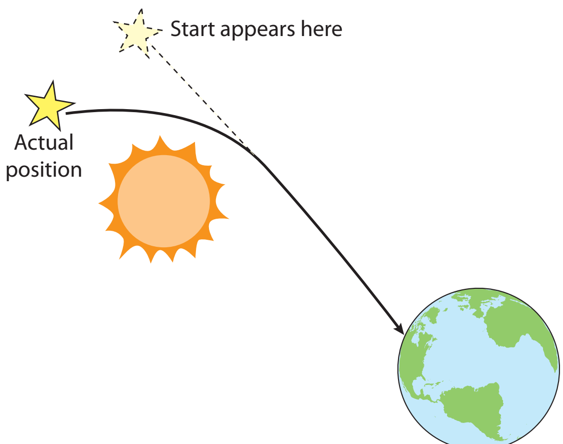


Figure 13.33 The light from a distant star is bent around the Sun. When viewed from the Earth, it appears to be to the right of its actual position.

Uncertainties

The uncertainties in Arthur Eddington's measurement of the angle of deflection were about the same size as the angle itself. This makes the results inconclusive. However, the results were hailed by the media as one of the most important discoveries of the time.

Gravitational lensing

If the Sun can bend light, then whole galaxies certainly will. When two galaxies are in line with an observer, the light from the far one will bend around the near one as in Figure 13.34. When observed through a powerful telescope, the light from the distant galaxy can be seen as a ring around the close one. This is called an Einstein ring, as shown in the photo.

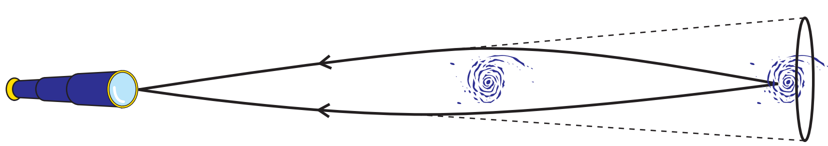


Figure 13.34 The formation of an Einstein ring by a large galaxy.

An Einstein ring seen with the Hubble space telescope. This is caused by gravitational lensing due to a large galaxy.

The slowing of time by large masses

If light has a constant velocity then it should always take the same time between two points. However, if a large mass is placed between the two points causing the light to follow a curved path, then it should take longer, that is, unless time slows down. To illustrate this point, consider driving from home to a friend's house in a car that travels always at 50 km h^{-1} along a straight road. You have arranged to meet at 12:30 so knowing the journey takes 30 minutes you leave at 12:00. Unfortunately there is a diversion and you have to drive a long way round, so you are going to be late.... unless, that is, you slow down your watch. Now, although you have driven further, the time taken will be the same. This doesn't work when visiting friends unless the diversion happens to be around a black hole.

Clocks near large masses tick slowly.

Gravitational red shift

If you were to view a clock ticking next to a massive distant galaxy you would notice that it was ticking more slowly than your own clock. A pendulum oscillating next to the clock would therefore oscillate more slowly than an identical one next to your clock. This would be true for all oscillating objects, so a source of EM radiation next to the galaxy would produce radiation that had a lower frequency, therefore longer wavelength than an identical source next to you. This effect is called gravitational red shift. Of course, an observer next to the galaxy using their own clock would measure the frequency to be normal, so if the wavelength that you receive is different it must have changed on the way. What happens is that as the radiation leaves the massive body, it loses energy. The energy of a photon is given by $E = hf$, so a loss of energy results in a reduced frequency. The Earth's gravitational field is big enough to cause a measurable difference between two sources placed at the top and bottom of a high building. This experiment was first performed at Harvard University in 1960 by Pound and Rebka. In this experiment gamma rays emitted from the nuclear decay of iron-57 at the top of a tower were compared with gamma rays from an identical source at the bottom. They found that the gamma rays from the top had a higher frequency than the ones at the bottom. This is in agreement with general relativity, which predicted that clocks at the top of the tower would be faster than those at the bottom, since they are further away from the mass of the Earth.

The equation for the change of frequency Δf due to gravitation is

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

where
Δh = the difference in height
g = gravitational field strength
c = speed of light in a vacuum.

Exercises

18 The photons used in the Pound-Rebka experiment had 14.4 keV energy, and the building was 22.6 m high. Calculate the change in frequency between the top and the bottom.

13.6 Space–time and black holes

Assessment statements

H.7.5 Describe the concept of space–time.
H.7.6 State that moving objects follow the shortest path between two points in space–time.
H.7.7 Explain gravitational attraction in terms of the warping of space–time by matter.
H.7.8 Describe black holes.
H.7.9 Define the term Schwarzschild radius.
H.7.10 Calculate the Schwarzschild radius.
H.7.11 Solve problems involving time dilation close to a black hole.

Curvature of space–time

It is quite difficult to visualise curvature in three dimensions of space and time. To get a true understanding requires a mathematical model rather than a visual one. However, it is possible to get some idea by considering just two dimensions of space. A classical two-dimensional space world would be a flat plane as in Figure 13.35. Since there are only two dimensions, any movement you make must be on this plane; there is no such thing as up and down, so no forces perpendicular to the surface exist. Applying Newton's laws of motion to this world, bodies will move in a straight line unless acted upon by resultant force.

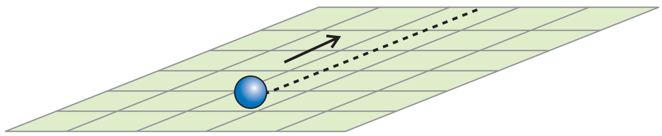


Figure 13.35 A blue ball with constant velocity travels in straight line.

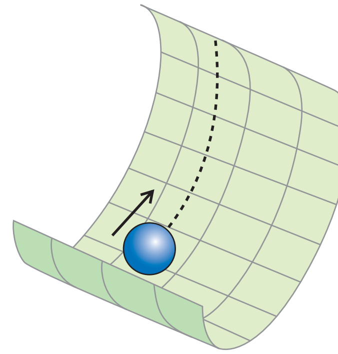


Figure 13.36 A ball moving with constant velocity in curved space has a curved path.

If we now curve space–time as in Figure 13.36, we find that a body travelling with constant velocity will travel in a curved path. You might think that it travels in a curved path because of the normal force from the plane, but remember this is a 2D world – no forces exist perpendicular to the plane.

Gravitational attraction between masses

According to Newtonian mechanics, all masses attract each other with a force that is proportional to their mass; this is supported by experiment. We all take for granted that gravity is the reason why things fall – but what is gravity? It is the thing that makes objects fall. Using curved space, we can explain how objects are attracted to each other. All masses curve space–time. The amount of curvature is related to the mass. So two masses placed close to each other will move towards each other due to the curvature, as illustrated in Figure 13.37. This shows how this might look in 2D curved space – each ball is rolling into the hole formed by the other.

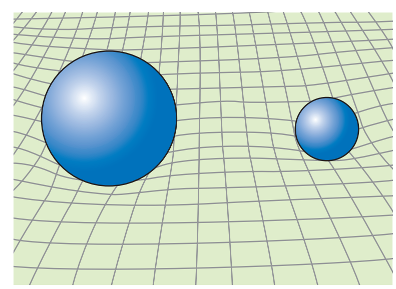


Figure 13.37 Each ball curves space–time.

Orbits

In the 2D model a large object curves space–time like a heavy man does when he sits on a bed. A large spherical mass put in the middle of the bed would form a bowl shape around it as in Figure 13.38. A ball pushed past the bowl with constant velocity would travel in a curved path around the mass. If it had just the right velocity it could go round and round the bowl like a golf ball does just before it goes down the hole; the ball is then in orbit around the large mass. Remember, no forces are needed if the ball is travelling in curved space.

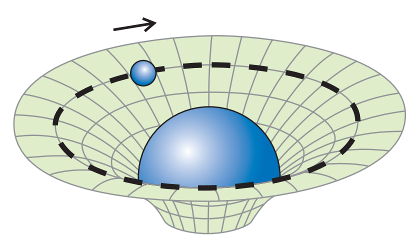


Figure 13.38 A small ball travels in a straight line in the curved space around a large mass.

Black holes

In Newtonian mechanics when a mass, m is lifted above the surface of the Earth it gains potential energy. This is explained in terms of the amount of work done (force $\times$ distance) to move the mass. If the field is uniform, this results in a value of mgh. When a body is thrown upwards, it therefore loses KE and gains PE. But if the mass is zero, the gain in PE must be zero, so it should not lose energy. As we have seen, this is not the case in general relativity, where, due to the curvature of space–time, photons lose energy as they move away from the Earth. In the extreme case of a very large mass in a very small space, space–time is curved so much that light loses all of its energy and can't escape. Such an entity is called a *black hole*, since it would emit no light. Black holes are formed when a star collapses to a size that is small enough to prevent light escaping. Its radius at this moment is called the *Swarzschild radius*. The star doesn't stay this size, but keeps getting smaller until it becomes a point, trapping any light that is within this radius.

Using Newtonian mechanics, we derived an equation for the escape speed from a planet by equating the loss in KE to the gain in PE required for the body to reach infinity.

$$\tfrac{1}{2}mv^2 = \frac{GMm}{R}$$

This gave a value for the escape speed, v from a planet of mass M and radius R to be

$$v = \sqrt{\frac{2GM}{R}}$$

This shows us that the escape speed will be great when the planet has a large mass and small radius. If we apply this equation to a collapsing star of known mass, we can find the radius, R_s at which it will become a black hole.

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

Rearranging gives

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

This equation, derived using Newtonian mechanics, happens to agree with the prediction of general relativity. However, we know that only particles with zero mass can travel at the speed of light (photons) so if we put this mass into the first line of the derivation we get $0 = 0$!

Event horizon
Once a body gets closer to a black hole than the Scharzschild radius, they would no longer be seen by an observer outside. It would be like someone disappearing over the horizon, and for this reason this line is called the *event horizon*.

Time dilation near a black hole

Close to a black hole, time slows down. In fact if you were to watch someone fall into a black hole, their motion would get slower and slower until they stopped just as they got to the edge of the event horizon, where they would remain frozen in time for ever. To calculate the extent of this time dilation, we can use the formula

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

Where Δt_0 is the time for the event as measured by the observer next to the black hole and Δt is the length of time it would take in your frame of reference a long way from the black hole.
R_S = the Scharzschild radius
r = the distance of the unfortunate observer from the centre of the black hole.

Exercises

19 A star of mass 2×10^{31} kg collapses to a point forming a black hole. What is the Schwarzschild radius of the black hole?

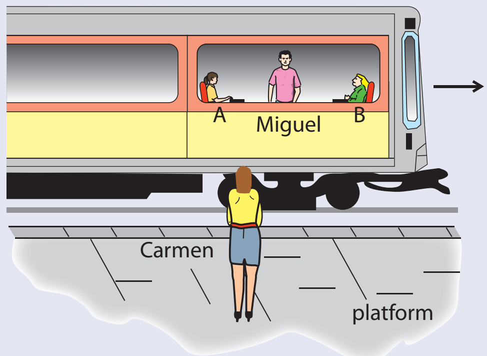
20 A spaceship with a light flashing once every minute travels towards the black hole in Question 19. An observer on the Earth (thankfully a long way from the black hole) watches the spaceship. Calculate time between flashes of the light when the spaceship is
(a) 10^6 km from centre of the black hole
(b) 100 km from the centre
(c) 60 km from the centre.

Practice questions

1 This question is based upon a thought experiment first proposed by Einstein.

(a) Define the terms *proper time* and *proper length*. (2)

In the diagram opposite Miguel is in a railway carriage that is travelling in a straight line with uniform speed relative to Carmen who is standing on the platform. Miguel is midway between two people sitting at opposite ends A and B of the carriage.



At the moment that Miguel and Carmen are directly opposite each other, the person at end A of the carriage strikes a match as does the person at end B of the carriage. According to Miguel these two events take place simultaneously.

(b) **(i)** Discuss whether the two events will appear to be simultaneous to Carmen. (4)

(ii) Miguel measures the distance between A and B to be 20.0 m. However, Carmen measures this distance to be 10.0 m. Determine the speed of the carriage relative to Carmen. (2)

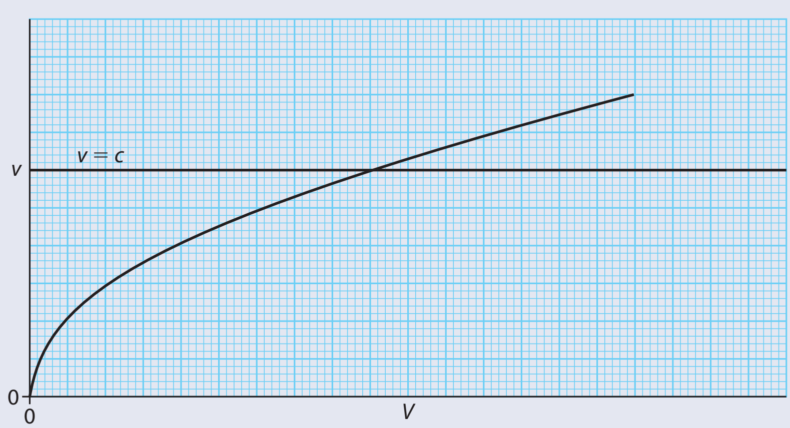
(iii) Explain which of the **two** observers, if either, measures the correct distance between A and B? (2)

(Total 10 marks)

2 This question is about electrons travelling at relativistic speeds.

A beam of electrons is accelerated in a vacuum through a potential difference *V*.

The sketch-graph below shows how the speed *v* of the electrons, as determined by non-relativistic mechanics, varies with the potential *V*, (relative to the laboratory). The speed of light *c* is shown for reference.



(a) On the grid above, draw a graph to show how the speed of the electrons varies over the same range of *V* as determined by relativistic mechanics.
(Note this is a sketch-graph; you do not need to add any values) (2)

(b) Explain briefly, the general shape of the graph that you have drawn. (3)

(c) When electrons are accelerated through a potential difference of 1.50×10^6 V, they attain a speed of 0.97 *c* relative to the laboratory.

Determine, for an accelerated electron,

(i) its mass. (3)

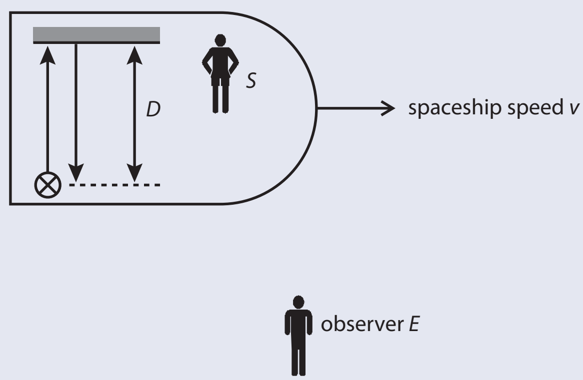
(ii) its total energy. (2)

(Total 10 marks)

3 This question is about time dilation.

(a) State what is meant by an *inertial* frame of reference. (1)

An observer *S* in a spacecraft sees a flash of light. The light is reflected from a mirror, distance *D* from the flash, and returns to the source of the flash as illustrated below. The speed of light is *c*.



(b) Write down an expression, in terms of *D* and *c*, for the time T_0 for the flash of light to return to its original position, as measured by the observer *S* who is at rest relative to the spaceship. (1)

The spaceship is moving at speed *v* relative to the observer labelled *E* in the diagram. The speed of light is *c*.

(c) **(i)** Draw the path of the light as seen by observer *E*. Label the position F from where the light starts and the position *R* where the light returns to the source of the flash. (1)

(ii) The time taken for the light to travel from *F* to *R*, as measured by observer *E*, is *T*. Write down an expression, in terms of the speed *v* of the spacecraft and *T*, for the distance *FR*. (1)

(iii) Using your answer in (ii), determine, in terms of *v*, *T* and *D*, the length *L* of the path of light as seen by observer *E*. (2)

(iv) Hence derive an expression for *T* in terms of T_0, *v* and *c*. (4)

(Total 10 marks)

4. This question is about relativistic motion.

The radioactive decay of a nucleus of actinium-228 involves the release of a β-particle that has a **total energy** of 2.51 MeV as measured in the laboratory frame of reference. This total energy is significantly larger than the **rest mass energy** of a β-particle.

(a) Explain the difference between *total energy* and *rest mass energy*. (2)

(b) Deduce that the Lorentz factor, as measured in the laboratory reference frame, for the β-particle in this decay is 4.91. (3)

A detector is placed 37 cm from the actinium source, as measured in the laboratory reference frame.

(c) Calculate, for the laboratory reference frame,

(i) the speed of the β-particle. (2)

(ii) the time taken for the β-particle to reach the detector. (2)

The events described in **(c)** can be described in the β-particle's frame of reference.

(d) For this frame,

(i) identify the moving object. (1)

(ii) state the speed of the moving object. (1)

(iii) calculate the distance travelled by the moving object. (2)

(Total 13 marks)

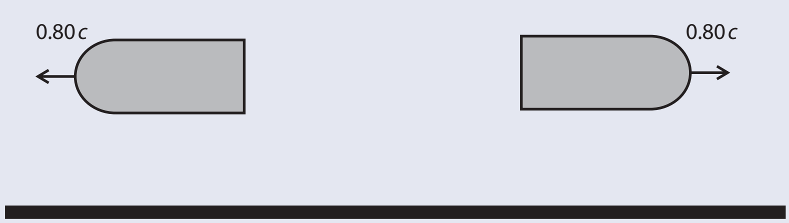
5. This question is about the postulates of special relativity.

(a) State the **two** postulates of the special theory of relativity.

Postulate 1

Postulate 2 (2)

(b) Two identical spacecraft are moving in opposite directions each with a speed of 0.80 c as measured by an observer at rest relative to the ground. The observer on the ground measures the **separation** of the spacecraft as increasing at a rate of 1.60 c.



ground

(i) Explain how this observation is consistent with the theory of special relativity. (1)

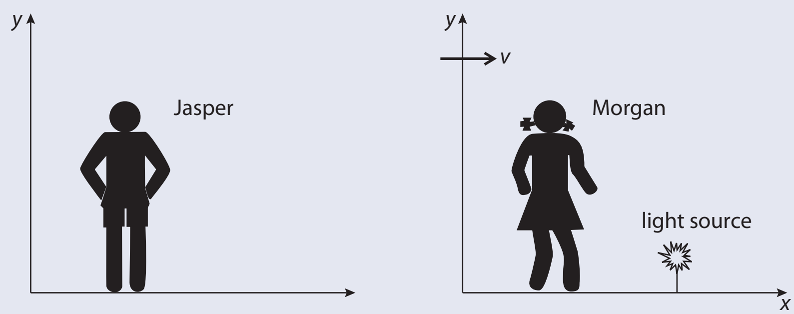
(ii) Calculate the speed of one spacecraft relative to an observer in the other. (3)

(Total 6 marks)

6. This question is about frames of reference.

(a) Explain what is meant by a *reference frame*. (2)

In the diagram below, Jasper regards his reference frame to be at rest and Morgan's reference frame to be moving away from him with constant speed v in the x-direction.



Morgan carries out an experiment to measure the speed of light from a source which is at rest in her reference frame. The value of the speed that she obtains is c.

(b) Applying a Galilean transformation to the situation, state the value that Jasper would be expected to obtain for the speed of light from the source. (1)

(c) State the value that Jasper would be expected to obtain for the speed of light from the source based on Maxwell's theory of electromagnetic radiation. (1)

(d) Deduce, using the relativistic equation for the addition of velocities, that Jasper will in fact obtain a value for the velocity of light from the source consistent with that predicted by the Maxwell theory. (3)

In Morgan's experiment to measure the speed of light she uses a spark as the light source. According to her, the spark lasts for a time interval of 1.5 μs. In this particular situation, the time duration of the spark as measured by Morgan is known in the Special Theory of Relativity as the proper time.

(e) **(i)** Explain what is meant by *proper time.* (1)

(ii) According to Jasper, the spark lasts for a time interval of 3.0 μs. Calculate the relative velocity between Jasper and Morgan. (3)

(Total 11 marks)

14 Medical physics (Option I)

14.1 The ear and hearing

Assessment statements

I.1.1 Describe the basic structure of the human ear.
I.1.2 State and explain how sound pressure variations in air are changed into larger pressure variations in the cochlear fluid.
I.1.3 State the range of audible frequencies experienced by a person with normal hearing.
I.1.4 State and explain that a change in observed loudness is the response of the ear to a change in intensity.
I.1.5 State and explain that there is a logarithmic response of the ear to intensity.
I.1.6 Define intensity and also intensity level (IL).
I.1.7 State the approximate magnitude of the intensity level at which discomfort is experienced by a person with normal hearing.
I.1.8 Solve problems involving intensity levels.
I.1.9 Describe the effects on hearing of short-term and long-term exposure to noise.
I.1.10 Analyse and give a simple interpretation of graphs where IL is plotted against the logarithm of frequency for normal and for defective hearing.

It is thought that the first creatures to live on Earth lived in water so their ears were adapted to receive sounds from the water. The sound waves would therefore travel directly from the water into the fluid in the ear. As animals moved to the land their hearing would not have been as good, since a lot of sound would be reflected as the sound waves pass from the air to the fluid in the ear. The development of our ears through years of evolution has been driven by this problem.

Although the ear is normally studied in biology classes, its operation is based on physical principles, and it is these we shall focus on here.

The function of the ear is to convert sound into electrical signals that can be processed by the brain. The outer ear is filled with air, so when a sound wave arrives at the ear it passes through the auditory canal. At the end of this canal is a thin membrane called the *eardrum*. The changing pressure in the air next to this membrane causes it to move in and out, which disturbs a small bone called the *mallus* (hammer). These vibrations are then passed on via two more small bones, the *incus* (anvil) and *stapes* (stirrup), to another thin membrane stretched across a window on the *cochlea*, a shell-like bone filled with fluid. As the stapes pushes this membrane in and out, the fluid inside the cochlea moves back and forth, disturbing small hairs. These small hairs are attached to cells that convert the movement into an electrical signal, which is sent to the brain along the auditory nerve.

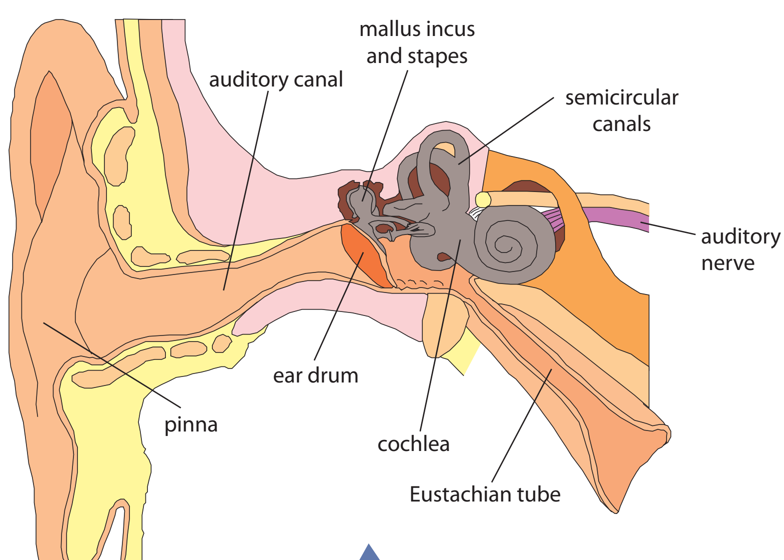


Figure 14.1 The structure of the ear.

Townsend's big eared bat.

Sound waves

Sound is produced by a vibrating object, such as a loudspeaker cone. As the cone moves back, the pressure of the layer of air in front of the speaker is reduced; this causes air in the next layer to flow towards the speaker leaving a reduced pressure further out. In this way the change of pressure propagates. If the cone vibrates sinusoidally, then a sinusoidally varying change in pressure will be propagated. A sound wave is therefore a propagation of a disturbance of pressure.

If you observe the layers of air, they move back and forth in the direction of the wave; this means it is a longitudinal wave. However, don't forget that individual air molecules are moving in random motion.

The outer ear

As sound waves approach the ear, they will be reflected from the sides and channelled into the auditory canal. Figure 14.2 shows a simplified version of this, rather similar to the old-fashioned hearing aid. Animals such as bats that rely on sound more than we do, have much larger, more cone-shaped ears that can be moved to point in the direction of the sound.

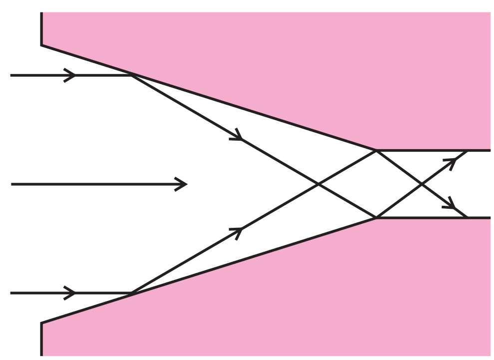


Figure 14.2 Sound waves are channelled by the outer ear.

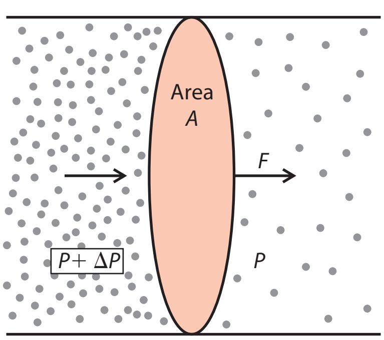


Figure 14.3 The sound wave causes an unbalanced force on the eardrum.

The eardrum

The eardrum is a thin membrane situated at the end of the auditory canal. When the sound wave reaches this point, the pressure of the air next to the eardrum will change. The air on the inside of the eardrum is kept at constant pressure due to the Eustachian tube connecting it to the back of the nose, so this change in pressure results in an unbalanced force that causes the eardrum to move as illustrated in Figure 14.3.

The force pushing from the left side of the eardrum $= (P + \Delta P)A$ and the force from the right $= PA$ so the unbalanced force is ΔPA.

The force exerted on the eardrum is not enough to move the liquid in the cochlea so it needs to be amplified; this is done in the next part of the ear.

Loud noises
To protect the inner ear from loud noises, the ear responds by tightening the eardrum.

The middle ear

Amplification in the middle ear takes place in two different ways; firstly the small bones (ossicles) that pass on the vibration are arranged as levers. These increase the force in the same way that a screwdriver increases the force you apply when you use it to get the lid off a tin of paint. The way the ossicles move is rather complicated but the principle is the same as illustrated in Figure 14.4

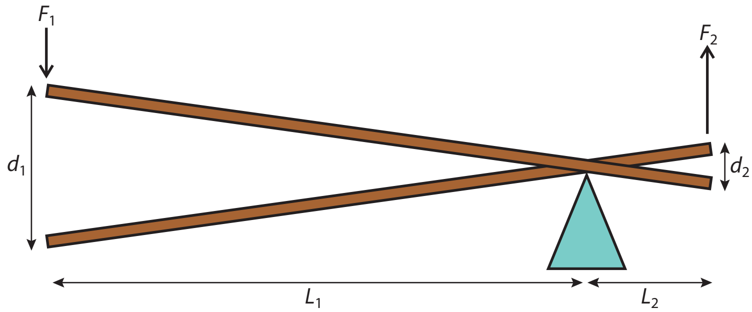
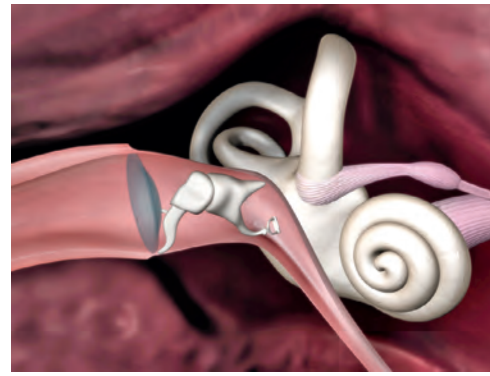


Figure 14.4 A force F_1 is applied to the left hand side of a lever which causes a force F_2 on the right hand side.



Ossicles in the middle ear.

The work done by $F_1 = F_1 d_1$
and work done by $F_2 = F_2 d_2$

Since energy is conserved, $F_1 d_1 = F_2 d_2$

so $\dfrac{F_1}{F_2} = \dfrac{d_2}{d_1}$

Since d_2 is smaller than d_1 then F_2 must be bigger than F_1.

In this way the force is made bigger. The amplification factor is equal to $\dfrac{d_1}{d_2}$.
Since the triangles are similar, this equals $\dfrac{L_1}{L_2}$.
Amplification factor $= \dfrac{L_1}{L_2}$

The second amplification occurs because the window to the cochlea is smaller than the eardrum. A given force will therefore result in a higher pressure. This is illustrated in Figure 14.5. To simplify matters, the lever action of the ossicles has been omitted.

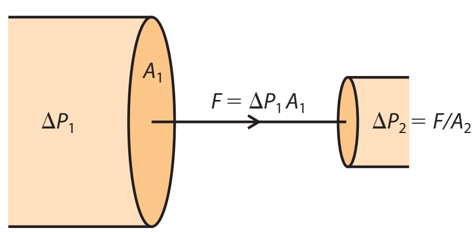


Figure 14.5 The force on the eardrum is passed on to the cochlea.

Increase in pressure ΔP_1 causes a force $F = \Delta P_1 A_1$ on the eardrum. This force is passed onto the cochlea via the ossicles. At the cochlea the force is applied to a window of area A_2 so the increase in pressure is $\dfrac{F}{A_2} = \dfrac{\Delta P_1 A_1}{A_2}$.
Since A_1 is bigger than A_2, the increase in pressure in the cochlea is greater than it was in the air.

Amplification factor $= \dfrac{A_1}{A_2}$

The inner ear

The cochlea is basically a 2 cm long coiled tube containing a liquid. At one end of the tube there is the oval window covered by a thin membrane (this is what the last of the ossicles vibrate against). At the other end is another membrane-covered opening called the round window. When the stapes pushes the oval window the

Damping
The ossicles don't only increase the force from the eardrum, they also damp its vibration. If the hammer were not resting on it, the eardrum would continue to vibrate long after the sound had stopped (like a real drum).

Impedance
When a sound travels from air to a liquid, 99.5% of it is reflected. This is due to the different impedance of the media. The amplification by the ossicles raises the amount of sound transferred to 50%.

Young people and loud music
When young people listen to loud music the ear responds by moving the stirrup away from the oval window reducing the pressure on the cochlea. This ability is lost with age.

Damage to the inner ear
If the movement of the fluid in the cochlea is too big, the small hairs will be destroyed, resulting in permanent reduction in hearing.

fluid is pushed along the tube causing the round window to bulge outwards. As the stapes vibrates, a sound wave is sent along the tube disturbing small hairs. It would be nice if we could explain, using simple physics, how we perceive the disturbance of these hairs as sound, but neither life nor the ear is that simple, so the rest of the story is best left to the biologists.

Exercises

The area of the eardrum is about 40 mm^2. A sound wave causes the air pressure to change by 0.2 Pa.

1 Calculate the unbalanced force on the eardrum.

2 The lever action of the ossicles amplifies the force experienced by the eardrum by a factor of 1.5. Referring to Figure 14.4, what is the ratio $\frac{L_1}{L_2}$?

3 What force will be transmitted to the oval window?

4 If the oval window has an area of 3 mm^2, calculate the change in pressure of the fluid in the cochlea.

5 Calculate the total amplification.

Frequency and pitch

Frequency is the number of complete cycles of a wave passing a point per unit time and is measured in hertz (Hz). The way we perceive sounds of different frequency is by their pitch; high pitch is high frequency.

If no one is listening to a sound does it still have a pitch?

The normal range of frequency that a human can hear is from 20 to 20 000 Hz. This range becomes smaller as people age, which is why you can buy high pitch ring tones for your mobile that your teacher can't hear (assuming your teacher is over 25).

Intensity and loudness

The intensity of a sound is the power delivered per unit area (the area is perpendicular to the direction of sound). A point source of sound produces a spherical wavefront. So, if the power generated by the source is P then the intensity at a distance r from the

source, $I = \frac{P}{4\pi r^2}$

The intensity is therefore proportional to $\frac{1}{r^2}$

If your ear has an area A then the power received at a distance r is

$$P_r = \frac{P}{4\pi r^2} \times A = IA$$

Loudness is the way we perceive intensity; a loud sound has a high intensity. They are not, however, linearly related; if you double the intensity the sound isn't twice as loud.

If you were to make a scale from the quietest sound at 0 to the loudest you can withstand at 10 then it may look something like Table 1. Try thinking of sounds that would fit in the gaps.

Table 1

0	Silence
1	Pin drop
2	
3	
4	
5	Talking
6	
7	
8	
9	
10	Rock concert

If we now measure the intensity of these sounds, we find that as we go up one point on the scale the intensity increases by a factor of 10, so the scale is logarithmic, as illustrated in Figure 14.6.

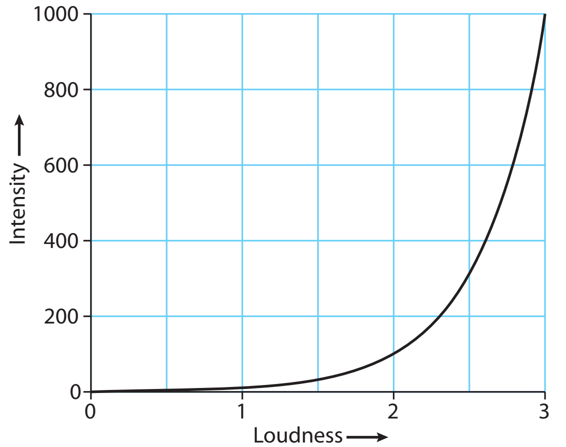


Figure 14.6 Graph of loudness against intensity. Only a small section is shown due to the large *y* scale.

So, although according to our loudness scale, a rock concert is 5 units louder than talking quietly, the intensity of the sound is 10^5 times bigger. Another way of looking at this is to think how many people talking would you need to make the same loudness as a rock concert? If the loudness scale were linear, the answer would be 5, but since it's logarithmic the answer is 10^5 (although they would all have to be standing on the same spot).

Sound intensity level (IL)

Since loudness is what we perceive, it can't really be quantified; instead we refer to the sound intensity level. This is a logarithmic scale so is more closely related to the loudness than the intensity.

The equation for this relationship is

$$IL\,(\text{bels}) = \log_{10}\frac{I}{I_0}$$

where I_0 is the threshold of hearing (the quietest noise detectable).

For normal ears $I_0 = 10^{-12}\,\text{W m}^{-2}$ which is equivalent to the sound made by a mosquito at a distance of 5 m.

Perception
A ticking clock may not sound loud in your room during the day but can stop you getting to sleep in the middle of the night.

The decibel scale dB

The decibel is $\frac{1}{10}$ bel. This makes a more convenient scale with 100 divisions from the quietest to the loudest noise. A 1 dB change is just noticeable.

$$\text{Loudness (decibels)} = 10\log_{10}\frac{I}{I_0}$$

Table 2 The decibel scale

0	Mosquito at 5 m
20	Ticking watch by your ear
50	Talking quietly
80	Vacuum cleaner at 1 m
90	Next to a busy road
100	MP3 player on full
110	Front of a rock concert
130	Jet taking off from 100 m (painful)
160	Eardrum bursts (very painful)

Worked examples

1 What is the sound intensity of the sound from a vacuum cleaner at a distance of 1 m?

2 What is the sound intensity level from two vacuum cleaners at 1 m?

Solution

1 From the table, the intensity level of the sound from a vacuum cleaner is 80 dB, so

$$80\text{ dB} = 10\log_{10}\frac{I}{I_0}$$

therefore

$$I = I_0 \times 10^8$$

$$I_0 = 10^{-12}\text{ W m}^{-2}$$

so $I = 10^{-4}\text{ W m}^{-2}$

2 From Example 1, we know that the intensity of one vacuum cleaner is 10^{-4} W m^{-2} so the intensity of two vacuum cleaners will be 2×10^{-4} W m^{-2}

$$\text{IL} = 10\log_{10}\frac{2 \times 10^{-4}}{1 \times 10^{-12}}$$

$$\text{IL} = 83\text{ dB}$$

Adding sounds
When there are two sounds at the same time you can add the intensity of the sounds but not their sound intensity levels.

Loudspeaker ratings
The power rating of a loudspeaker is the maximum power that can be delivered to it, not the power of the sound that comes out. A lot of power is lost due to heat.

Exercises

6 A loudspeaker gives out 1 mW of sound. Assuming the wavefront spreads out in a sphere, calculate:
(a) the intensity of the sound at a distance of 1 m
(b) the sound intensity level at 1 m.

7 A person standing 50 m from the stage at a rock concert hears a sound of 100 dB. If they move to 10 m from the stage:
(a) what was the intensity at 50 m?
(b) what will the intensity of the sound be at 10 m?
(c) how many dB will they hear?

8 To produce a sound of 100 dB at a distance of 25 m, how many speakers each giving out 1 mW would you need?

9 A good home stereo speaker will give 100 dB at a distance of 1 m when 1 W of power is fed to it.
(a) What is the intensity of the sound?
(b) How much power is given out by the speaker?

The effects of exposure to loud sounds

Short-term exposure to sounds over 100 dB can damage the hairs in the cochlea, leading to tinnitus or ringing in the ears. As the length of time one is exposed to a sound increases, the loudness also increases, as does the likelihood of sustaining damage. This is why people who work with everyday machinery like lawnmowers, which are not particularly loud, wear ear protectors. Long exposure to sounds above 90 dB can lead to permanent hearing loss as the hair cells in the cochlea begin to die. Pete Townsend, the lead guitarist in The Who, once the loudest band in the world, developed severely reduced hearing, probably due to prolonged use of headphones with the volume too high.

Hearing test

The ear is not equally sensitive to all frequencies, so in order to test someone's hearing, a range of different frequencies must be used. During a hearing test, sounds of frequency 125, 250, 500, 1000, 2000, 4000 and 8000 Hz are played through headphones. Each frequency starts so quietly it cannot be heard and the intensity is gradually increased until the patient hears a sound. This intensity level (the hearing level) is recorded. The results are then plotted on a graph of hearing level in dB against frequency, as shown in Figure 14.7. This is called an *audiogram*.

Audiograms are made for both ears and also with a vibrator attached to bone behind the ear; this sends sound waves directly to the cochlea through the bone, so will reveal whether the problem is in the middle/outer ear or cochlea.

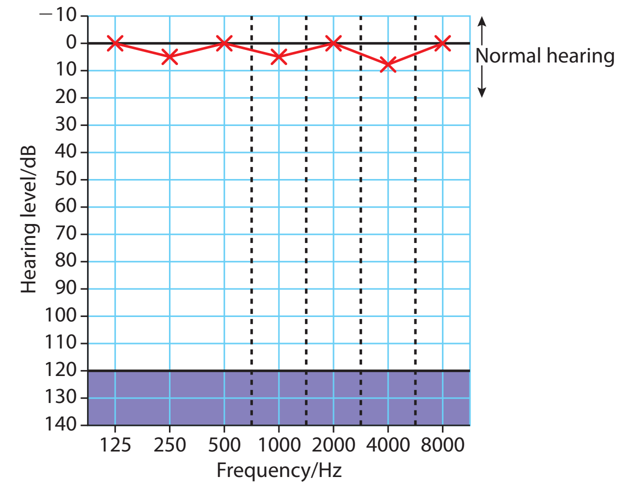


Figure 14.7 The audiogram of someone with good hearing. Note that the frequency scale is logarithmic.

Interpreting audiograms

In the example shown in Figure 14.8, the hearing through the bone is fine, but the hearing through the air is poor. This means that the cochlea is functioning properly, so the problem must be in the outer or middle ear. Maybe the auditory canal is blocked, the eardrum is damaged or the ossicles are not working properly. This can be corrected by using a hearing aid that increases the intensity of the sound coming into the ear (conductive hearing loss).

Example 1

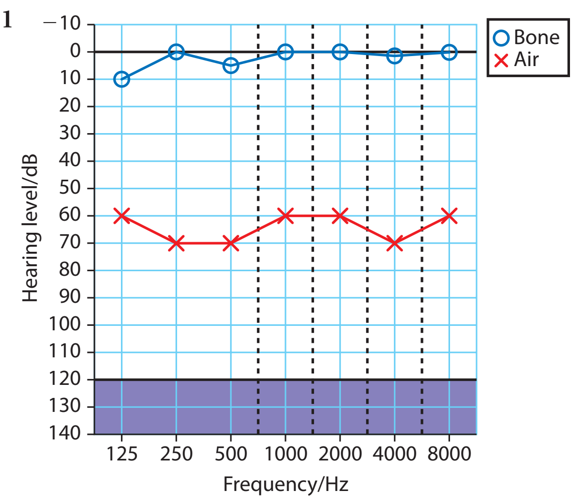


Figure 14.8 Audiogram 1.

In Figure 14.9, both the hearing through the air and the bone is poor so the problem probably lies in the cochlea. This cannot be corrected by a regular hearing aid but can be improved by a cochlea implant. This works by sending electrical impulses from a microphone fixed above the ear directly to the nerves in the cochlea (sensory hearing loss).

Prolonged exposure to loud noise will lead to a reduction in hearing around 4 kHz. This shows up as a pronounced dip in the audiogram.

Example 2

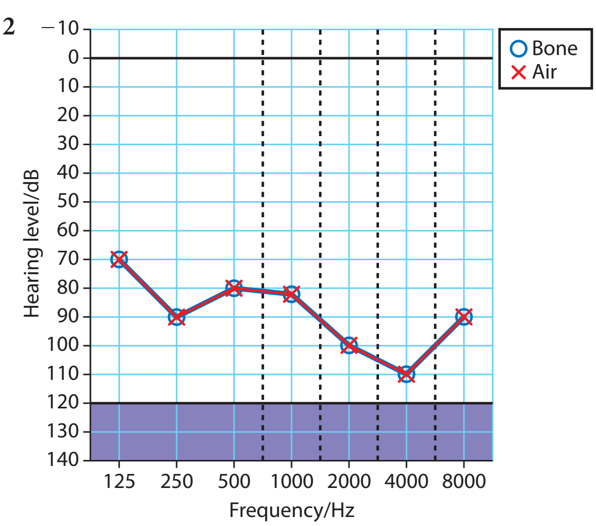


Figure 14.9 Audiogram 2.

Wearing ear protectors at work.

Noise and the law

Many people work with noisy machinery and can suffer damage to their hearing as a result of their work. In many countries employers are required by law to ensure that the sound levels are reduced, or to supply all workers with personal ear protectors. In addition, workers who are at risk (work areas with high levels of noise above 85 dB) should be given regular hearing tests. People who choose to enter a noisy area, for example, a rock concert are not protected by the law, but people who work there *are* protected.

The regulations that protect hearing are different from country to country. See if you can find out what the regulations are in your country.

14.2 Medical imaging

X-rays

Assessment statements

I.2.1 Define the terms attenuation coefficient and half-value thickness.
I.2.2 Derive the relation between attenuation coefficient and half-value thickness.
I.2.3 Solve problems using the equation $I = I_0e^{-\mu x}$
I.2.4 Describe X-ray detection, recording and display techniques.
I.2.5 Explain standard X-ray imaging techniques used in medicine.
I.2.6 Outline the principles of computed tomography (CT).

X-rays pass through human tissue in the same way as light passes through glass, but they do not pass through bone as easily. X-rays affect photographic film in the same way as light, so if a part of the body is exposed to X-rays from above, a reverse shadow picture of the bones can be obtained on a photographic plate placed underneath as in the photo.

The author's repaired ankle. The light parts are the shadows.

Absorption of X-rays

Light passing through a solid is absorbed if the energy of the photon is sufficient to increase the energy of an electron such that it can change from a low energy level to a high one. Since visible light photons have an energy which is about the same as the energy of an atomic electron (about 1 eV) most are absorbed. X-ray photons have a lot more energy (>120 eV) so when they interact with atomic electrons, they knock the electron out of the atom in one of two ways:

Photoelectric effect:

The X-ray photon is absorbed completely giving all of its energy to the KE of the electron.

Compton scattering:

The photon gives some of its energy to the electron and continues with less energy and therefore longer wavelength $\left(E = \frac{hc}{\lambda}\right)$

Exponential equation
As the number of photons gets less the number absorbed per centimetre gets less. This can be written as a differential equation:

$$\frac{-dN}{dx} \propto N$$

The solution of this equation is of the form

$$N = Ae^{-kx}$$

This is very similar mathematically to nuclear decay.

Attenuation

When a parallel beam of X-ray photons pass through a solid, some of them interact with the atoms of the solid. This leads to a reduction in intensity, or *attenuation*. The number of interactions between the photons and solid depends on how many photons there are, so, as the beam progresses through the solid, the number of interactions decreases. This results in an exponential relationship between the beam intensity and the distance travelled through the absorber as illustrated by the graph in Figure 14.10.

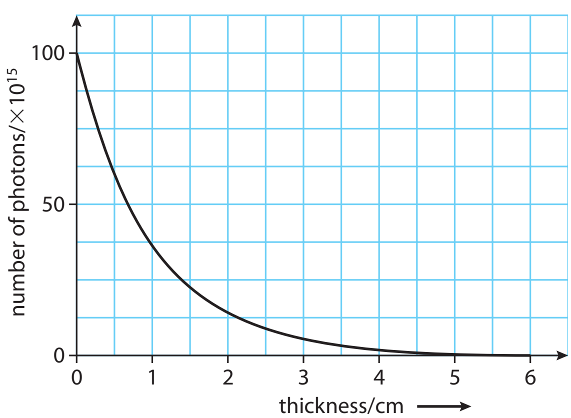


Figure 14.10 Graph of number of photons (proportional to intensity) against absorber thickness.

Since the intensity is proportional to the number of photons, we can deduce that the intensity of the radiation, I after it has passed through a thickness of absorber, x is given by the equation:

$$I = I_0 e^{-\mu x}$$

Where I_0 = the original intensity

μ = the attenuation coefficient

Half value thickness ($x_{\frac{1}{2}}$)

The half value thickness is the thickness required to reduce the intensity to one half the original.

If the thickness = $x_{\frac{1}{2}}$ then

$$I = \frac{I_0}{2} = I_0 e^{-\mu x_{\frac{1}{2}}}$$

Taking logs gives

$$x_{\frac{1}{2}} = \frac{ln2}{\mu} = \frac{0.693}{\mu}$$

Half-value thickness and energy
The half-value thickness of X-rays depends on their energy. High energy X-rays have a larger value than low energy, so are more penetrating.

Worked example

A parallel beam of X-rays of intensity 0.2 kW m^{-2} is passed through 5 mm of a material of half value thickness 2 mm. Calculate the intensity of the beam.

Solution

First calculate the attenuation coefficient.

$$\mu = \frac{0.693}{x_{\frac{1}{2}}} = \frac{0.693}{2}\,\text{mm} = 0.35\,\text{mm}^{-1}$$

Now use the attenuation equation to find I

$$I = I_0 e^{-\mu x} = 0.2e^{-0.35 \times 5} = 0.035\,\text{kW m}^{-2}$$

Exercises

10 The intensity of a beam of X-rays is reduced from 0.1 kW m^{-2} to 0.08 kW m^{-2} after passing through 4 mm of a material. Calculate:

(a) the attenuation coefficient of the material

(b) the half-value thickness.

11 An X-ray beam of intensity 0.5 kW m^{-2} is passed through 3 mm of a material of half-value thickness 1 mm. Calculate

(a) the attenuation coefficient of the material

(b) the intensity of the beam passing through the material.

12 After passing through 6 mm of material, the intensity of an X-ray beam is reduced to 40%. Calculate:

(a) the attenuation coefficient

(b) the half-value thickness.

Taking an X-ray picture

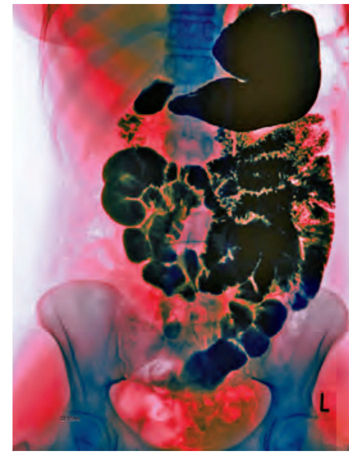
X-rays are often used to view broken bones. The simplest way to take an X-ray picture is to place the broken part on a piece of photographic paper and shine X-rays through it from above. The photo paper must be enclosed in a lightproof box, otherwise it will be darkened by the light. When the paper is developed, the areas that were exposed to the X-rays will be dark and the parts that the X-rays couldn't penetrate will be light. This is why the metal pins in the photo on page 490 are so light. To view breaks in bones, it is important to choose X-rays that aren't stopped by the bone too easily; if this is the case you would just see an outline of the foot. Small changes in bone thickness must cause significant changes in beam intensity, to give the picture contrast and making diagnosis possible. The problem is that X-rays are ionizing so can damage the cells of the body, so high energy X-rays and long exposure times can be dangerous. To overcome this problem, the photograph can be improved using an intensifying screen.

Intensifying screen

An intensifying screen is a sheet of fluorescent material that gives out visible light when the X-ray photons land on it. If this is placed on either side of the photo film then the light emitted causes the light areas to become lighter without increasing intensity or exposure time.

Barium meal

Bones show up well on an X-ray because small changes in thickness lead to a sufficient change in intensity. Soft tissue is not so easy to see, since the X-rays pass through even thick layers without much attenuation. To make soft tissue such as the gut show up more clearly on an X-ray photograph, it can be filled with an X-ray-opaque substance such as barium sulphate. In the case of the stomach, this is a simple matter of drinking a barium meal.



An X-ray of the stomach after a swallowing a barium meal.

Digital images

Modern X-ray machines are moving over to similar devices to those used in digital cameras. An array of photosensitive diodes gain charge when exposed to X-radiation. The charge creates a p.d. which can then be converted into a digital signal. The advantage of digital signals is that the image can be enhanced and even coloured electronically. Such systems are used in airport security, where luggage is scanned by a line of sensors as it passes along a conveyor belt.

Tomography

Sometimes a shadow picture isn't good enough, especially if the bit you are interested in is somewhere deep in the body obscured by other bits. In these cases it would be better if you could view a slice of the body; this is made possible with tomography. The way this works can be illustrated with light. Imagine two balls fixed in a transparent block of plastic as shown in Figure 14.11. A spotlight is shone on the red ball and moved to the right. As this happens the spotlight is kept focused on the red ball. The shadow of the red ball will therefore stay in the middle of the spotlight but the blue ball's shadow will move from the left to the right. If a piece of photographic film is now moved along with the spot of light, the shadow of the red ball will form a spot in the middle but the shadow of the blue ball will be a blurred line as in Figure 14.12.

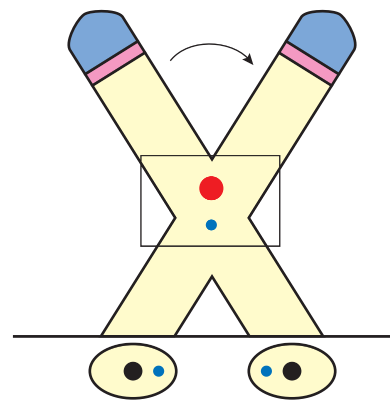


Figure 14.11 As the spotlight is moved, it is kept focused on the red ball. The yellow circles show the shadows.

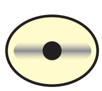


Figure 14.12 The photo of the shadow.

The same effect can be achieved by moving the X-ray source and photo film.

Computer tomography (CT scan)

Computer tomography is a more sophisticated version of tomography, where the X-ray source and a circular array of detectors are rotated around the patient. This does not give a picture directly but by analysing the signals from the detector with a computer, a 3D picture of a slice through the patient can be put together. By moving the detectors and X-ray source along the length of the patient's body, a complete 3D image of the patient can be built up. This can be digitally manipulated and artificially coloured to show different layers and highlight specific features.

A CT scan showing the skeleton plus the position of a transplanted kidney.

Ultrasound

Assessment statements

I.2.7 Describe the principles of the generation and the detection of ultrasound using piezoelectric crystals.
I.2.8 Define acoustic impedance as the product of the density of a substance and the speed of sound in that substance.
I.2.9 Solve problems involving acoustic impedance.
I.2.10 Outline the differences between A-scans and B-scans.
I.2.11 Identify factors that affect the choice of diagnostic frequency.

Safety
Since ultrasound is non-ionizing it can be used safely for long periods of time. This is also why it is used for scans of unborn babies.

Ultrasound is sound that has such a high frequency that we can't hear it, greater than 20 kHz. By analysing the ultrasound reflected off different layers in the body, it is possible to build up a picture of the internal structure.

The piezoelectric effect

When a quartz crystal is compressed or stretched a potential difference is induced across it; this is called the *piezoelectric effect*. This happens because the atoms in the quartz crystal are arranged in such a way so that when the crystal is deformed they become polarized, as illustrated in Figure 14.13.

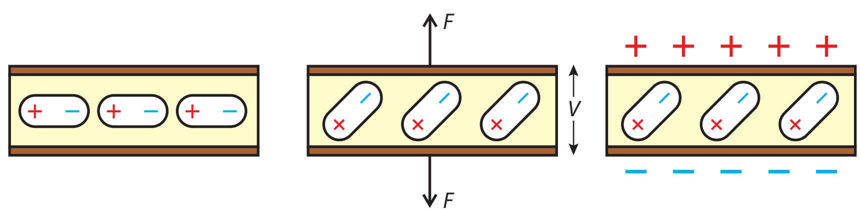


Figure 14.13 When the crystal is stretched the dipoles line up causing a p.d.. If a p.d. is applied the crystals line up causing expansion.

If a p.d. is applied across the crystal, the dipoles are made to line up with the electric field, resulting in expansion of the crystal. This is also shown in Figure 14.13. (Note: this is a simplification and does not show how compression would also cause a p.d., but this *does* happen.)

Ultrasound production and detection

To produce ultrasound, an alternating p.d. of frequency >20 kHz is applied to the crystal, causing it to vibrate. To detect ultrasound, we can use the alternating p.d. induced when a sound wave causes a crystal to vibrate. When performing an ultrasound scan, ultrasound reflections are analysed, so to detect the reflected wave, the detector and transmitter must be in the same place. If pulsed ultrasound is used the same crystal can be used for transmission and detection; however, it is important that the pulse is short enough so that the reflected wave doesn't come back before the transmitter has finished transmitting.

Example

If the pulse length is 10^{-6} s and the speed of sound in body tissue is 1500 m s^{-1} then the minimum distance that the wave can travel before the pulse has finished transmitting is $1500 \times 10^{-6} = 1.5$ mm. So it could be reflected off something at a depth of 0.75 mm.

Since all interesting organs are deeper than 1.5 to 2 mm, a pulse of 10^{-6} s duration would be fine.

Choice of frequency

If the ultrasound is diffracted, it spreads out and will not be reflected back to the detector. For this reason the wavelength must be short enough so that is not diffracted by anything that we wish to know the position of. The smallest objects that a doctor might be interested in are a few millimetres, so a wavelength a bit less than this would be suitable. If $\lambda = 1$ mm and the velocity $= 1500\,\text{m s}^{-1}$ then $f = 1500/10^{-3} = 1.5$ MHz. Higher frequencies would give better resolution but are absorbed more by the body, resulting in higher attenuation.

This means that the ultrasound operator has to juggle between pulse length and frequency to get the best possible image for organs of different depth.

Acoustic impedance

When ultrasound waves are incident on the boundary between two media, part of the wavefront is reflected and part refracted. The percentage reflected depends upon the relative acoustic impedance of the two media, where impedance is defined by the equation

$$\text{acoustic impedance, } Z = \rho c$$

where ρ = the density of the medium
c = the velocity of the ultrasound.

The unit of Z is $\text{kg m}^{-2}\,\text{s}^{-1}$.

The greater the difference in acoustic impedance, the greater is the percent reflection. The difference in impedance between the air and skin is very large, so to prevent almost all the ultrasound being reflected before it enters the body, the gap between the transmitter and skin is filled with a gel. This is normally smeared over the body or the probe before the scan begins.

Exercises

Material	Velocity of sound/ms^{-1}	Density/kgm^{-3}
Muscle	1540	1060
Bone	3780	1900
Fat	1480	900

13 Calculate the acoustic impedance of the different tissues in the table above.

14 Which pair of media will give the greatest percent reflection?

A-scans

The most basic way to display the data is to plot a graph of the strength of the reflected beam against time. From this it is possible to see the position of any changes in medium in the body directly in front of the probe, as illustrated in Figure 14.14.

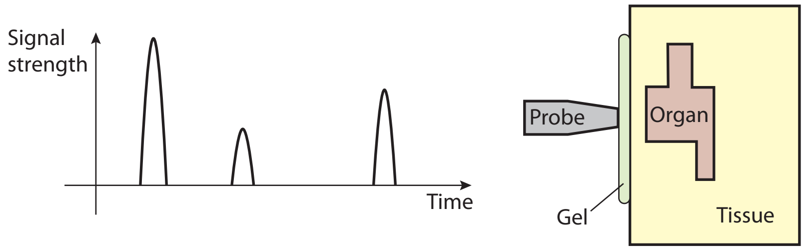


Figure 14.14 An A-scan for an organ surrounded by tissue. The depth and thickness of the organ can be deduced from the times of the reflected pulse. Notice how the second pulse is smaller than the first due to attenuation but the last is bigger – this is because most of the ultrasound is reflected off the last boundary.

To gain more information about the organ, the probe can be moved up and down; this will reveal the size, shape and changes in thickness.

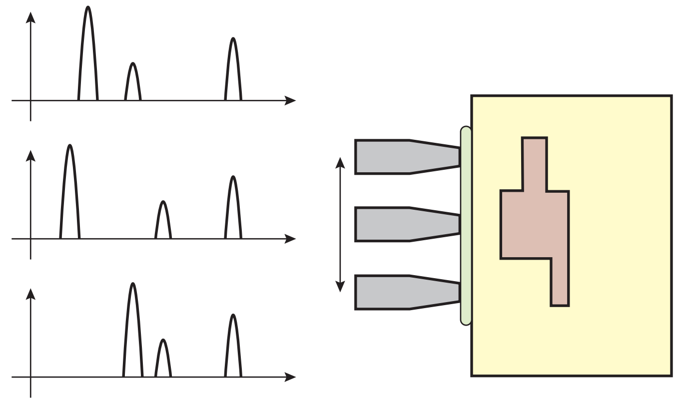
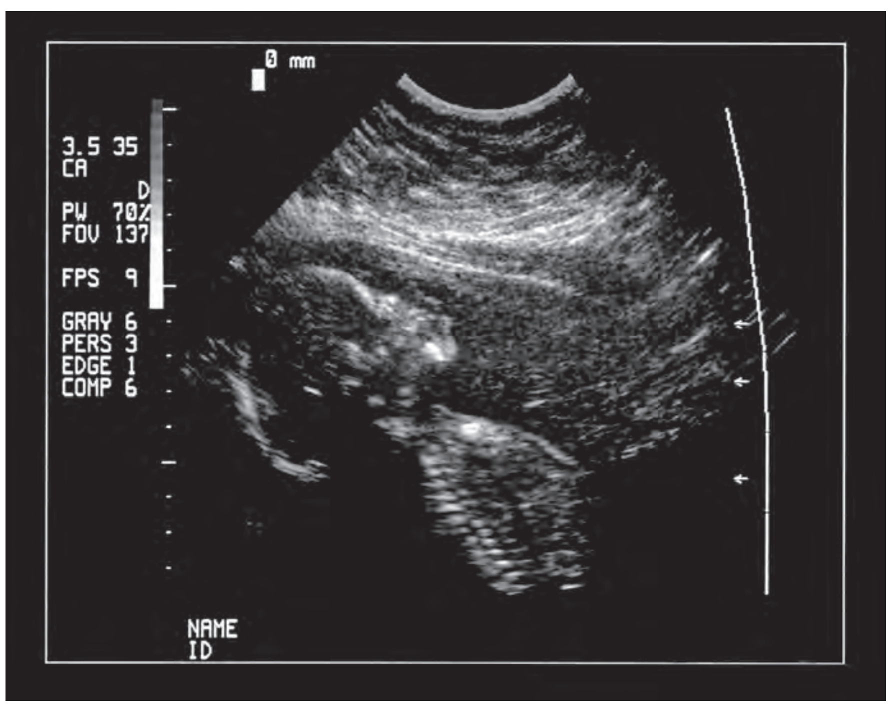


Figure 14.15 A-scans from three positions revealing the shape of the organ.

B-scans

An A-scan gives information, but it doesn't give an image of the organ. A B-scan converts the signal into a dot, the brightness of the dot corresponding to the strength of the signal. If the probe is swept across the organ this can be converted into an image.

Instead of sweeping the probe, a probe with many transmitters can be used.



Taking an ultrasound scan of an unborn baby. The baby's backbone and head are clearly visible in the image.

Exercises

15 The following questions relate to the A-scan shown in Figure 14.16. In this scan the ultrasound waves travelling at 1500 m s^{-1} reflect off both sides of an organ then off a bone. The time registered is the time from the pulse leaving the transmitter to returning to the detector.

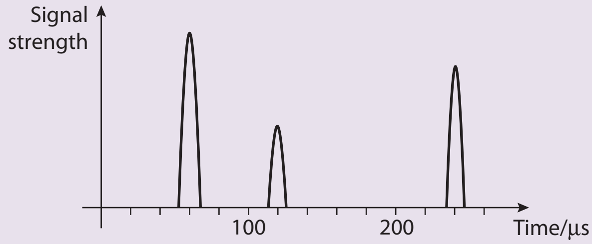


Figure 14.16

Calculate
(a) the depth of the organ
(b) the thickness of the organ.

3D ultrasound

By using an array of probes and moving them around the patient in three dimensions, a 3D image can be constructed using a computer.

A 3D scan of a foetus.

All these incredible ways of looking into the human body are immensely helpful in diagnosing illnesses. Unfortunately, the machines that create these images are very expensive and not available to all patients in all countries.

Nuclear magnetic resonance (NMR)

Assessment statements
I.2.12 Outline the basic principles of nuclear magnetic resonance (NMR) imaging.

Instead of sending waves into the body from an external source as in X-rays and ultrasound, NMR works by getting the hydrogen nuclei in the body to give out radio waves. By analysing this radiation it is possible to build a very detailed image.

Nuclei are magnetic

Due to a combination of the fact that the nucleus is charged and has spin, nuclei with odd numbers of protons or neutrons behave like magnets. This means that they will line up with an applied magnetic field in the same way as a compass needle. If a compass needle is pushed to one side, it will oscillate back and forth before eventually coming to rest pointing north again, and the same is true for nuclei.

Resonance

We know from studying SHM that if an oscillating system is forced to oscillate at its own natural frequency the amplitude of oscillation will be large; this is called *resonance*. If a hydrogen nucleus is placed in a strong magnetic field and displaced slightly, it will oscillate at a frequency of around 60 MHz. So if the nucleus is pushed at a frequency of 60 MHz it will oscillate with large amplitude; this can be done using radio waves. Once disturbed, the nucleus will continue to oscillate for several seconds. As it does this, it emits radio waves, which can be detected and used to build up an image.

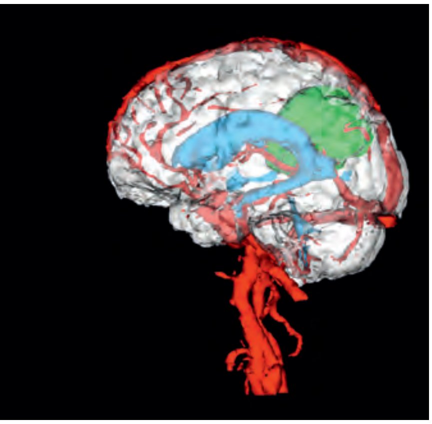
To see a simulation showing how MRI works, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 14.1.

Magnetic resonance imaging (MRI)

The radio waves emitted by the resonating nuclei are detected using a coil of wire and analysed with a computer. There are several aspects of the radiation that are useful for building the image.

Frequency: The frequency of the radiation depends upon the strength of the magnetic field (stronger magnet = higher frequency). If the magnetic field is made to vary from one place to another, then the position of the source can be found.

Relaxation time: The time taken for the oscillation to die away is different for different types of tissue. By measuring this time it is possible to determine the tissue type and add that into the image.



3D MRI of a brain showing a tumour in green.

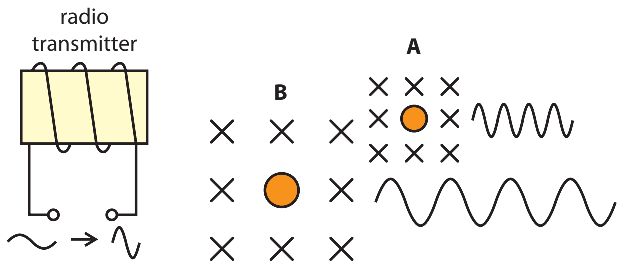


Figure 14.17 Nuclei in a region of strong field resonate at higher frequency, so if a high frequency signal is detected, it must have come from region **A**.

The use of lasers

Assessment statements

I.2.13 Describe examples of the use of lasers in clinical diagnosis and therapy.

Lasers aren't used to make images of organs, but are included here since they are another form of radiation that is useful to medicine. The most common use is as a scalpel (knife) used in surgery.

Cutting with a laser

Light from a light bulb spreads out in all directions, so when it is focused with a lens, only a small amount of the energy given out by the bulb will enter the lens. This light is not parallel so will not be focused on the same point, so a spot will be formed. Laser light, however, *is* parallel so doesn't spread out, and when focused by a lens, is brought to a focus at a point. This produces a very intense point of light.

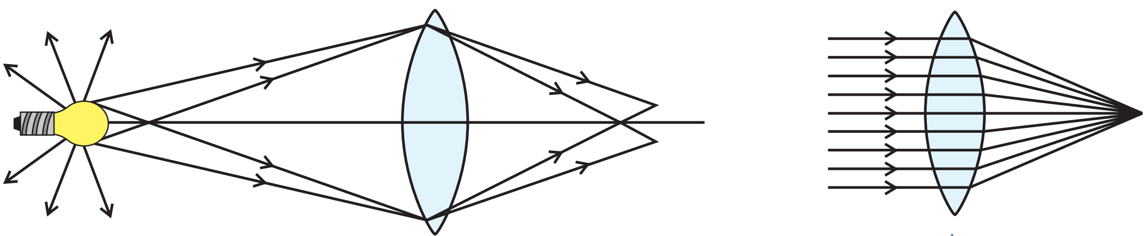


Figure 14.18 Focusing of light from a bulb compared to light from a laser.

In this way, the energy from the beam is focused on a very small amount of matter. With sufficient energy this piece of matter can be vaporized. Remember that the amount of energy required to change the state of an amount of matter is proportional to the mass, so if the mass is very small the amount of energy is not high.

The advantages of using a laser as a cutting tool in surgery is firstly that the laser can make a much finer cut, and secondly the vaporizing action seals small blood vessels, reducing bleeding.

Laser light can be passed through an optical fibre with a focusing device at the end; this enables surgeons to perform operations inside the body without opening it up. An example of this is laser angioplasty in which an optical fibre is fed along a blood vessel from the arm to the heart where it is used to burn away unwanted material.

The laser used in pulse oximetry is not focused or very powerful, so it doesn't burn a hole in your finger.

Measuring oxygen content of the blood

If you shine a flashlight into your mouth your cheeks glow red, due to the blood in your cheeks. This idea can be used to measure the amount of oxygen in the blood.

Since laser light is very intense it will shine through thin parts of the body like a finger or a baby's foot. A pulse oximeter contains two laser beams of different wavelength, one red the other infra red. Red light is absorbed most by blood with no oxygen and IR most by blood with oxygen. By measuring the relative absorption of the two wavelengths as they pass through the blood, it is possible to calculate the amount of oxygen in the blood, as well as monitoring the pulse. Without this device, nurses would have to take regular blood tests – time consuming for the nurse and uncomfortable for the patient.

A blood oxygen monitor (pulse oximeter).

14.3 Radiation in medicine

Assessment statements

I.3.1 State the meanings of the terms exposure, absorbed dose, quality factor (relative biological effectiveness) and dose equivalent as used in radiation dosimetry.
I.3.2 Discuss the precautions taken in situations involving different types of radiation.
I.3.3 Discuss the concept of balanced risk.
I.3.4 Distinguish between physical half-life, biological half-life and effective half-life.
I.3.5 Solve problems involving radiation dosimetry.
I.3.6 Outline the basis of radiation therapy for cancer.
I.3.7 Solve problems involving the choice of radioisotope suitable for a particular diagnostic or therapeutic application.
I.3.8 Solve problems involving particular diagnostic applications.

We have seen how radiation such as X-radiation can be used to make images of the inside of a human body because it can pass through solid matter. The images obtained are invaluable to a doctor who might be looking for the presence of a tumour or broken body part. Sometimes an illness is not caused by a broken part (that might be visible on an X-ray or scan) but by a malfunction, resulting in a failure of the organ to deliver the right fluid to the right place. Monitoring the movement of fluids around the body is difficult, but can be made possible by injecting a gamma ray emitting isotope into the fluid. Since gamma rays pass through solids, the path of the fluid can easily be followed from the outside, by measuring the movement of the gamma ray source. The problem is that ionising radiation, like gamma and X-rays, damage the cells of the body so great care must be taken to minimize this damage.

Safe dose
Exposure to any amount of ionising radiation is potentially harmful, so there is no such thing as a safe dose. In this way it is similar to smoking, except that exposure to radiation such as X-rays can have beneficial results, but smoking doesn't.

Radiation dosimetry

Radiation dosimetry is the calculation of how much radiation is absorbed as a result of exposure to different types of radiation, and the effect that this has on different parts of the body. This depends on the type of radiation (α, β, γ or X), the energy of the radiation and how much of it is actually absorbed.

Activity

The activity of a radioactive isotope is the number of disintegrations per second. This does not give much information about the effect it would have, since this depends on how ionizing the radiation is.

Background radiation
Every day our bodies are exposed to natural radiation from the Sun (cosmic rays), rocks, building materials and even food. The total dose equivalent for 1 year is 3 mSv.

Exposure (X)

Exposure is a measure of how much ionizing radiation you would be exposed to in a particular environment, based on the number of ions produced per kilogram of air. This quantity is only used for X or γ radiation since although α and β can produce a lot of ions, their range is so short that someone standing in the same room as the source but several metres away will not be exposed to any radiation.

Exposure, X is defined by the equation:

$$X = \frac{Q}{m}$$

where $Q =$ the total charge of all the positive ions produced
$m =$ the mass of air in the room.

The unit of exposure is $\mathrm{C\,kg^{-1}}$.

This gives some indication of the potential danger of entering this environment but doesn't give a true measure of the amount of radiation your body will absorb.

Absorbed dose (*D*)

The absorbed dose is a measure of the energy absorbed by the actual tissue. This is defined by the equation:

$$D = \frac{E}{m}$$

where $E =$ the total energy absorbed
$m =$ mass of tissue.

The unit of absorbed dose is $\mathrm{J\,kg^{-1}}$ or gray (Gy).

Dose equivalent (*H*)

Different types of radiation have different biological effects so to give a measurement of the biological effect of the radiation, the absorbed dose is multiplied by a factor that is dependent on the type of radiation. α is the most damaging, so the factor is 20, whereas for X, β and γ it is 1. This factor is called the *quality factor* (Q).

Dose equivalent is defined by the following equation:

$$H = QD$$

The unit of dose equivalent is also $\mathrm{J\,kg^{-1}}$. However, to distinguish between absorbed dose and dose equivalent, the unit sievert (Sv) is used. Since $Q = 1$ for X-rays, the dose equivalent for X-rays is the same as the absorbed dose. For other radiations, the dose equivalent gives the amount of X-radiation that would cause the same harm.

● **Hint:** The dose in Gy tells you how much energy per kg is absorbed by the body. The equivalent dose in Sv tells you how much harm it will do.

Radiation	Quality Factor
X	1
γ	1
β	1
α	20

Source	Dose/mSv
Background dose in 1 year	3
Ankle X-ray	0.02
CT scan of the head	2.0
Barium meal	8.0

Table 3 Examples of dose equivalents.

Dose/mSv	Effect
1000	Nausea, vomiting
2000	Loss of body hair
4000	Bleeding in the mouth
10 000	Death after 14 days
50 000	Death within 48 hours

Table 4 Effect of dose.

Exercises

16 A school gamma source emits 500 gamma ray photons per second, each with an energy of 1 MeV.

(a) How much energy in joules will be emitted in 1 hour?

(b) If 10% of the photons are absorbed by a student during an hour, how much energy will they absorb?

(c) If the mass of the student is 70 kg, what absorbed dose will they receive?

17 How much energy does a 70 kg person receive if they are exposed to 2 mSv of X-radiation?

18 How much energy does a 70 kg person receive if they are exposed to 2 mSv of alpha radiation?

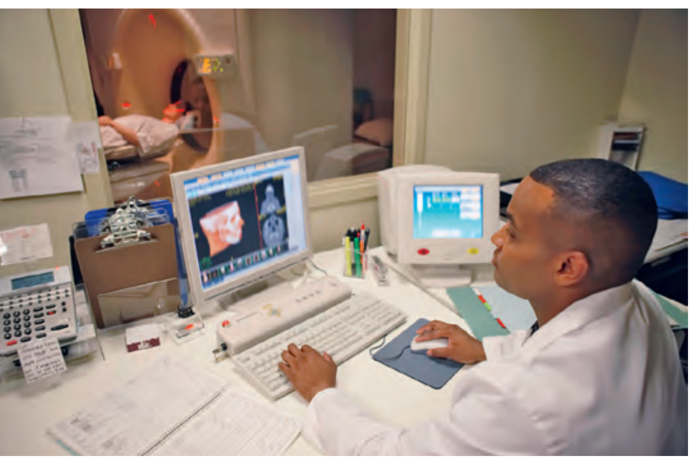
19 A 70 kg person receives an equivalent dose of 1S v from a source of gamma radiation in 1 hour:

(a) How much energy did the person absorb?

(b) How many 1 MeV gamma photons did they absorb?

(c) If they absorbed 10% of the photons from a source, how many photons came out of the source?

(d) What was the activity of the source?



The radiologist sits in a separate room.

Protection against radiation

There are three ways that exposure to radiation can be minimized:

- **Distance**. The intensity of all ionizing radiation decreases with distance, so increasing the distance between the source and person reduces the dose received.
- **Shielding**. Alpha radiation is stopped by paper, beta by a sheet of aluminium and gamma is reduced by a few centimetres of lead. By wearing protective clothing or standing behind a shield, the dose can be reduced.
- **Time**. The dose received from a radioactive source is proportional to the time spent exposed to the radiation, so reducing this time will reduce the dose.

If you have ever had an X-ray you will have seen these three preventative methods in action. When the radiologist switches on the X-ray machine they will probably be sitting in a separate room behind a shield. You can't avoid the X-radiation but each time you have an X-ray, the time of exposure will be kept to a minimum and the dose received will be monitored to make sure you don't exceed the maximum yearly dose.

Monitoring of radioactive dose

People who work with ionizing radiation such as radiologists, nuclear energy workers and particle physicists, must be careful to keep their exposure to a minimum. This can be done by monitoring the amount of radiation in the workplace, but this only gives the exposure not the absorbed dose. To give a better measure of the amount of radiation absorbed by an individual, each worker wears a personal detector such as a film badge. Photographic paper is a cheap and effective way to detect ionizing radiation. When an ionizing particle is absorbed by the paper, it causes a chemical change in one of millions of tiny grains that coat it. When the paper is processed the grains turn black, and the number of these black grains is an indication of the amount of radiation received. To distinguish between different types and energies of radiation the paper is separated into different regions, each region has a filter of different thickness in front of it. Since the film must be processed, the badge has to be handed in at regular intervals (e.g. once a month). If a worker is found to be over the limit they will have to have a medical check and probably time off work.

International limits

As previously mentioned, there is no safe dose of radiation so the limits set by bodies such as the International Commission on Radiological Protection are based on what is considered acceptable. It's not ethically acceptable to experiment with people to find out how much radiation is safe – however, there is a lot of data that was gathered from the survivors of the atomic explosions at Hiroshima and Nagasaki. From this data it is possible to find out what level of radiation is certainly not acceptable, then it's a matter of working backwards to decide what is acceptable.

Is it problematic to use the data collected from the victims of Hiroshima and Nagasaki?

The internationally accepted limit for people working with radioactive materials is 50 mSv per year, but for the general public it is 5 mSv per year.

Balanced risk

All exposure to ionising radiation is harmful but sometimes the benefits outweigh the possible harm. The table below shows some examples, so you can decide if the benefit outweighs the risk.

The rapidly dividing cells in a fetus are particularly susceptible to damage by radiation; this is why pregnant women should avoid contact with sources of radiation.

Risk	Benefit
Ankle X-ray (0.02 mSv)	Ankle gets repaired properly
Radiation emitted by smoke detector	Detector might detect a fire
CT scan of unborn baby to find out if it's a boy or girl (2 mSv)	Know whether to buy blue or pink booties.

When weighing up the risk versus benefit of exposure to radiation the 'as low as reasonably achieved' (ALAR) principle is usually applied.

Radiation therapy

Side effects
The side effects of radiation therapy are the same as radiation sickness: loss of hair, nausea etc.

The operation of a living cell is dependent on many interlinked chemical reactions. Ionization changes the chemical properties of atoms, so when the atoms of a cell are ionized, some of the chemical reactions are altered. Cells have developed to be able to repair themselves if the damage is minor (if this were not the case bac kground radiation would be killing cells all the time) but if too many atoms are affected, or if the affected atoms are particularly important, the cell will die. This is not normally a good thing unless the cell happens to be one that you would like to remove, such as a cancer cell. In this case exposure to radiation can be beneficial.

External irradiation

One way to kill cancer cells is to irradiate them with a beam of gamma radiation. This will also affect healthy cells, but since cancer cells are not properly functioning cells, they are unable to repair themselves when exposed to levels of radiation that other cells can withstand. To specifically target the cancer cells the beam is passed through the body from different directions that intersect at the site of the cancer; in this way the cancer will get the highest dose. This method can only be used when the cancer is localized.

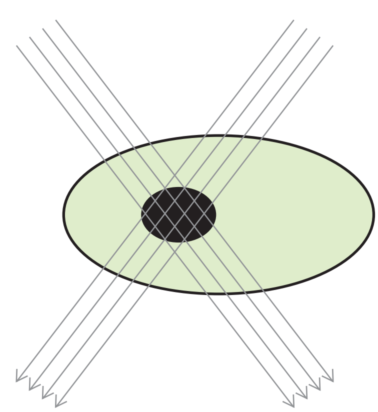


Figure 14.19 The γ ray beams cross at the tumour (black).

Internal irradiation

The body can be irradiated from the inside either by placing a solid radioactive source next to the tumour (*brachytherapy*) or by injecting or ingesting a fluid containing the radioactive isotope. The chemical properties of all the isotopes of an element are the same, so, for example, you can have carbon dioxide that contains carbon-12 or carbon-14. If radioactive isotopes are introduced into the bloodstream, they will follow the same path that the substance would normally follow. In this way, radioactive iodine can be used to treat cancers of the thyroid, since if you eat iodine it collects there.

Choice of isotope

Once inside the body, the radioactive isotope must deliver enough radiation to kill the cancer cells but not the patient. Effectiveness of the treatment depends upon several factors:

- **Energy**. High energy radiation will have a more damaging effect (higher dose equivalent) on the cancer cells but will also be more penetrating, so may pass through the cancer into neighbouring healthy cells.
- **Type**. Alpha radiation is most damaging but not very penetrating, so would need to be applied very close to the cancer cells. Beta and gamma are more penetrating but not so ionizing, so would need longer exposure times and would have more effect on surrounding tissue.
- **Chemical properties**. If a particular organ is to be targeted using an ingested fluid then the isotope used must be an element that collects at that place, for example, iodine in the thyroid. If the isotope is in solid form to be placed next to the cancer then it must not react in a harmful way with the surrounding tissue.
- **The radioactive half-life**. If the isotope is in the form of a solid pellet or wire, then the half-life is not so important since it can be removed when its job is done. If anything, a long half-life is best so it can be used many times. Caesium-137 is an example of an isotope used as a solid. This has a half-life of about 30 years. For ingested isotopes the half-life should be long enough so a small amount can deliver a large enough dose, but short enough so the patient does not remain radioactive for a long time after the treatment.
- **The biological half-life**. Even after a long time, a radioactive isotope will retain some level of activity so the best choice of isotope is one that the body gets rid of naturally (in sweat, urine etc). The rate at which a substance is removed from the body is proportional to the amount in the body, so has an exponential decay similar to radioactivity. We can't therefore say how long it will be in the body but can quote how long it takes for half to be removed. Water, for example, has a biological half-life of 10 days, but lead in the bone has a biological half-life of 10 years.

Effective half-life

The time taken for the amount of a substance to be reduced by half (T_E) is a combination of the physical half-life (T_p) and the biological half-life (T_B) and can be found using the equation:

$$\frac{1}{T_E} = \frac{1}{T_p} + \frac{1}{T_B}$$

Example

Iodine-131 is gamma emitter with a half-life of 8 days that accumulates in the thyroid. This makes it suitable for the treatment of thyroid cancers. The biological half-life of iodine in an adult is 80 days, so the effective half-life, T_E is given by:

$$\frac{1}{T_E} = \frac{1}{8} + \frac{1}{80}$$

Effective half life = 7.3 days

Radioactive tracers

If a small amount of a radioactive isotope is injected into the blood, its journey as it flows through the blood vessels can be tracked by measuring the spread of radioactivity around the body with a GM tube. Furthermore, since activity is proportional to the amount $\left(\frac{dN}{dt} = -\lambda N\right)$ it is possible to find out how much substance there is by simply measuring the activity. This makes it possible to monitor the distribution of the isotope around the body. This technique is called *radioactive tagging* and has many uses in medicine.

Measuring blood volume

To measure the volume of water in a large tank, a small amount of radioactive isotope can be mixed with 1 litre of water and its activity measured. The water plus isotope is then put back in the tank and mixed up, distributing the isotope around the tank. A second sample of 1 litre is now taken out of the tank and its activity measured. This will be smaller than the activity of the first sample, since the isotope concentration is now much less.

The ratio of original activity/diluted activity will give the volume in litres. The same technique can be used to measure blood volume using small amounts of iodine-131.

$$\frac{\text{Original activity}}{\text{Diluted activity}} = \frac{\text{total volume}}{\text{sample volume}}$$

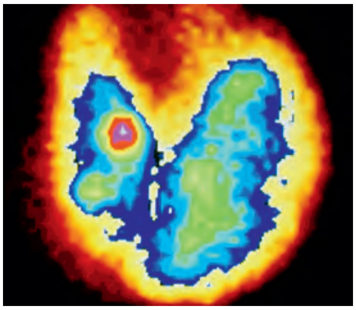
Thyroid activity

The function of the thyroid is to make chemical messengers (hormones) used to control the operation of other organs. One of the elements needed to make the hormones is iodine, which the thyroid collects from the blood. Iodine flows in and out of a healthy thyroid at a constant rate; any malfunction will result in an accumulation. By using radioactive iodine as a tracer, the flow of iodine through the thyroid can be monitored.

Calcium build-up in heart muscle

The build-up of calcium is an indication that heart muscle is damaged. If radioactive technetium is injected into the bloodstream, it will take the place of calcium in the heart muscle, giving an indication of the amount of calcium and hence damage to the muscle.

This thyroid scan shows that the right side is enlarged and not functioning properly, the red area on the left is an area of extra activity caused by the left side having to compensate.

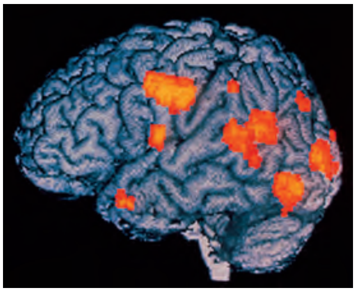


Imaging using tracers

The radiation emitted from a radioactive tracer can be used to produce an image of the organ in which it collects. The photo shows the picture of a thyroid taken with a camera sensitive to gamma radiation six hours after the patient had eaten some iodine-131.

PET scan

In Positron Electron Tomography (PET), a carbon-11 tracer is introduced into the body when the patient breathes in some carbon monoxide. The carbon monoxide attaches itself to the red blood cells and is carried around the body. Carbon-11 decays by the emission of positrons which annihilate with electrons to produce two gamma ray photons travelling in opposite directions. The gamma radiation is then detected and used to construct an image of the brain where the areas of most activity (more blood) can be identified.



A PET scan of the brain of a schizophrenic man hallucinating about coloured heads speaking. The scan shows activity in the areas of the brain responsible for vision and hearing.

Exercises

20 Calculate the effective half-life of technetium-99 which has a physical half-life of 0.25 days and a biological half-life of 1 day.

21 10 cm^3 of fluid containing iodine-131 with an activity of 100 Bq is injected into the blood. Some time later a 10 cm^3 sample of blood is taken and found to have an activity of 0.2 Bq. What is the volume of the blood? (Ignore the decay of iodine-131.)

Practice questions

1 This question is about sound and hearing.

(a) State the approximate range of frequencies that are audible to a person with normal hearing. (1)

(b) Outline the mechanism by which different frequencies are distinguished in the cochlea. (3)

(c) A person with normal hearing can just hear a sound of intensity 10^{-12} W m^{-2} at a frequency of 1000 Hz.

(i) A sound wave of frequency 1000 Hz incident on the ear drum has an intensity of 2.7×10^{-5} W m^{-2}. Calculate the sound intensity level at the ear. (2)

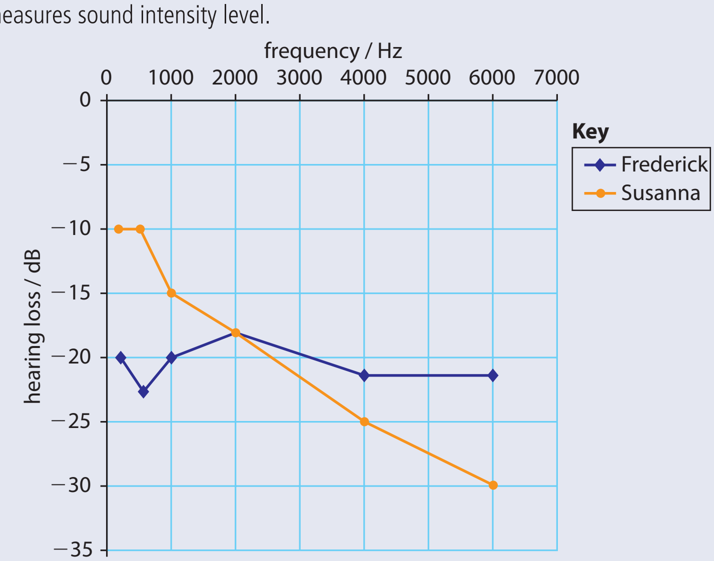
(ii) Explain why the response of the ear is measured as a change in sound intensity level rather than a change of intensity of sound. (3)

(Total 9 marks)

2 This question is about hearing loss and audiograms.

(a) Distinguish between conductive and sensory hearing loss. (2)

The diagram below shows the audiograms for two people, Frederick and Susanna, both of whom are suffering from hearing loss. The hearing loss is measured in decibels, a unit that measures sound intensity level.



(b) Outline how sound intensity level is related to sound intensity. (2)

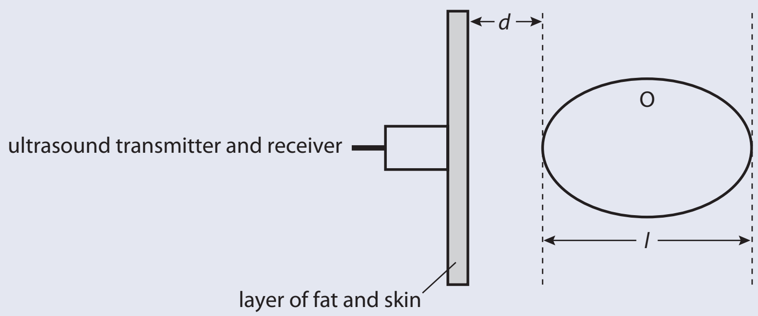
(c) Suggest the type of hearing loss from which each person could be suffering and state a possible cause of the hearing loss. (4)

(Total 8 marks)

3 This question is about ultrasound scanning.

(a) State a typical value for the frequency of ultrasound used in medical scanning. (1)

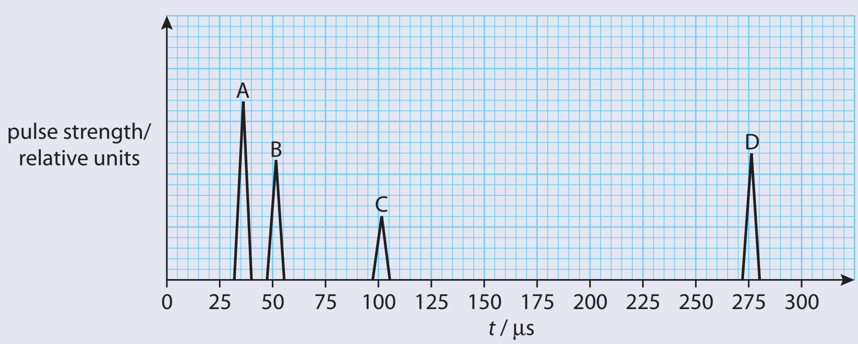
The diagram below shows an ultrasound transmitter and receiver placed in contact with the skin.



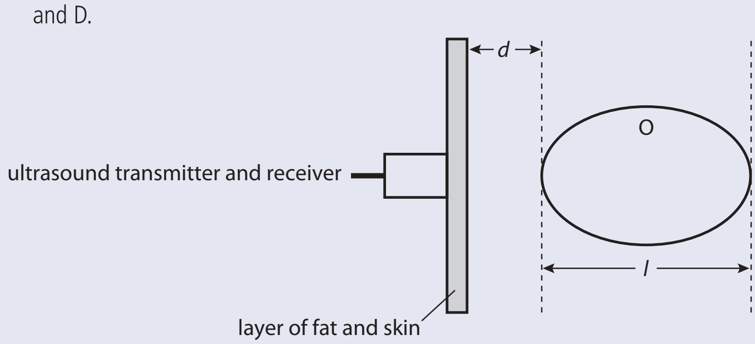
The purpose of this particular scan is to find the depth *d* of the organ labelled O below the skin and also to find its length, *l*.

(b) **(i)** Suggest why a layer of gel is applied between the ultrasound transmitter/receiver and the skin. (2)

On the graph below the pulse strength of the reflected pulses is plotted against time *t* where *t* is the time lapsed between the pulse being transmitted and the time that the pulse is received.



(ii) Indicate on the diagram below the origin of the reflected pulses A, B and C and D.



(2)

(iii) The mean speed in tissue and muscle of the ultrasound used in this scan is $1.5 \times 10^3\,\mathrm{m\,s^{-1}}$. Using data from the above graph, estimate the depth *d* of the organ beneath the skin and the length *l* of the organ O. (4)

(c) The above scan is known as an A-scan. State **one** way in which a B-scan differs from an A-scan. (1)

(d) State **one** advantage and **one** disadvantage of using ultrasound as opposed to using X-rays in medical diagnosis. (2)

(Total 12 marks)

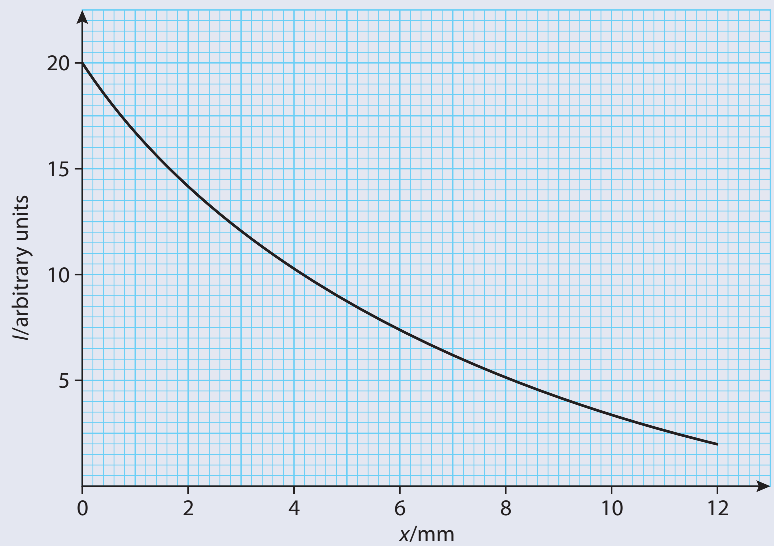
4 This question is about medical imaging.

(a) State and explain which imaging technique is normally used

(i) to detect a broken bone. (2)

(ii) to examine the growth of a fetus. (2)

The graph below shows the variation of the intensity *I* of a parallel beam of X-rays after it has been transmitted through a thickness *x* of lead.



(b) **(i)** Define *half-value thickness*, $x_{\frac{1}{2}}$. (2)

(ii) Use the graph to estimate $x_{\frac{1}{2}}$ for this beam in lead. (2)

(iii) Determine the thickness of lead required to reduce the intensity transmitted to 20% of its initial value. (2)

(iv) A second metal has a half-value thickness $x_{\frac{1}{2}}$ for this radiation of 8 mm. Calculate what thickness of this metal is required to reduce the intensity of the transmitted beam by 80%. (3)

(Total 13 marks)

5 This question is about the biological effectiveness of radiation.

(a) Explain the term *quality factor (relative biological effectiveness)*. (1)

(b) A beam of protons is directed at a tumour of mass 0.015 kg. In order to kill the tumour, a dose equivalent of 240 J kg^{-1} is required.

Using the data below, determine the exposure time, assuming all the incident protons are absorbed within the tumour.

Energy of each proton = 4.0 MeV.

Number of protons incident on the tumour per second = 1.8×10^{10}.

Quality factor for protons of this energy = 14. (4)

(Total 5 marks)

6 This question is about radioactive isotopes of iodine.

The isotope iodine-131 is used to treat malignant growths in the thyroid gland. The isotope has a physical half-life of 8 days and a biological half-life of 21 days.

(a) Explain the term *biological half-life*. (2)

(b) Calculate the effective half-life of the isotope. (2)

The isotope iodine-123 has a physical half-life of 13 hours.

(c) Suggest why it is preferable to use this isotope for **imaging** the thyroid rather than iodine-131 (1)

(Total 5 marks)

15 Particles (Option J)

15.1 Description and classification of particles

Assessment statements

J.1.1 State what is meant by an elementary particle.
J.1.2 Identify elementary particles.
J.1.3 Describe particles in terms of mass and various quantum numbers.
J.1.4 Classify particles according to spin.
J.1.5 State what is meant by an antiparticle.
J.1.6 State the Pauli exclusion principle.

Setting the scene

Modern particle physics can be said to have begun in the early 1930s. Before we can understand the development of particle physics, we must know something about what particles had been discovered and what theories had been proposed to describe their interactions at this time.

A very brief (and selective) history of discoveries

1898 Joseph Thomson measured the charge/mass ratio of the electron, leading to the 'plum pudding' model of the atom.

1905 Albert Einstein explained the photoelectric effect by proposing that light is made of photons.

1911 Ernest Rutherford suggested that the atom consists of a nucleus with surrounding electrons.

1931 James Chadwick found evidence for the neutron.

1930 Wolfgang Pauli explained why beta particles have a spread of energies, by proposing the existence of the neutrino.

1931 Paul Dirac proposed that every particle has an antiparticle with opposite charge.

In the 1930s physicists were developing the theories of quantum mechanics.

Table 1 Particles and antiparticles.

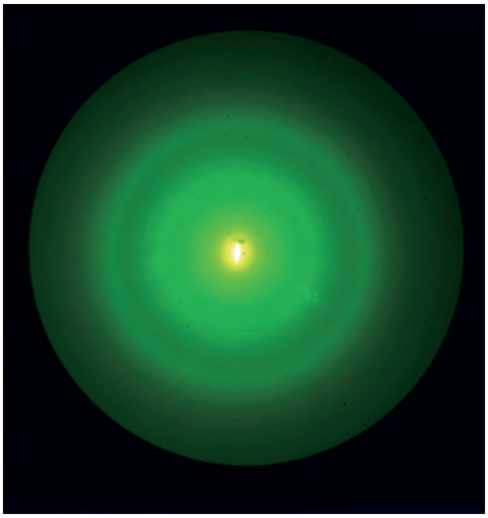
Particle	Antiparticle
electron	positron
proton	antiproton
neutron	antineutron
neutrino	antineutrino

Why do we think an electron is a particle?

In physics, when we talk about a particle, we mean a small ball. When we say an electron is a particle we don't mean it *is* a small ball we mean it *has the properties* of a small ball. What are those properties? If a ball is thrown forwards, it falls in a parabolic path due to the gravitational field of the Earth. When an electron is projected forwards in a uniform electric field, it also follows a parabolic path. When balls hit each other, momentum and energy are conserved. When two electrons hit each other, momentum and energy are conserved. So we say electrons have particle-like properties.

Particle nature of light

Light is electromagnetic radiation which has the wave-like properties of reflection, refraction, diffraction and interference. However, light is produced when an electron changes energy level and this results in a packet or quantum of energy. These packets are called *photons*. A photon has particle-like properties, for example it can bounce off an electron, resulting in a change of momentum.



Electron diffraction: the pattern produced when electrons pass through a powder made of crystals.

Wave nature of particles

We have seen that electrons have particle-like properties, but they also have wave-like properties. If electrons are passed through a crystal they land in a pattern that is very similar to the pattern formed when light passes through a diffraction grating. The electrons are being diffracted. The electrons are behaving like waves, the wave giving the probability of finding the electron.

Probability waves

If one ball is travelling towards a stationary ball and we know the velocity of the moving one, then we can calculate exactly when they will hit. When two electrons travel towards each other, we cannot predict exactly when they will hit, but we can say when they are *most likely* to hit. We cannot write an exact (x, y, z) coordinate to give the position of an electron, but we can say where it is *most likely* to be. This leads to a probability distribution, and the equation for that probability distribution is similar to the equation for a wave. This is also the case for all sub-atomic particles.

Heisenberg's uncertainty principle

Since the position of a particle is defined by a probability distribution, it is not possible to define exactly where it is. Werner Heisenberg developed the idea that you cannot precisely know the position and momentum of a particle at the same time. To know where something is you must shine light on it, but if you do that you will change its momentum. This can be expresses in the equation:

$$\Delta p \Delta x \geqslant \frac{h}{4\pi} \qquad \left(\frac{h}{4\pi} = 5.28 \times 10^{-35}\,\mathrm{J\,s^{-1}}\right)$$

where Δp = uncertainty in momemtum

Δx = uncertainty in position

h = Planck's constant

An alternative way of writing this is:

$$\Delta E \Delta t \geqslant \frac{h}{4\pi}$$

where ΔE = uncertainty in energy
Δt = uncertainty in time

It is important to realize that this is not just about measurement; this is the way things are defined. This means that the energy of a particle is not an exact value but within some spread of values ΔE. The shorter the time, the bigger ΔE can be. This does not agree with the classical view that energy is always the same.

Spin
Particles are said to have spin. This doesn't mean that they are spinning but they do have some similar properties to a spinning ball.

Particle	Spin
electron	$\frac{1}{2}$
proton	$\frac{1}{2}$
neutron	$\frac{1}{2}$
photon	1

Table 2

Pauli exclusion principle

The Pauli exclusion principle is fundamental to the way particles interact; it says that two or more electrons cannot occupy the same quantum state at the same time. This means that two electrons cannot exist in the same place, so when two electrons approach each other they will repel each other, unlike photons of light that pass through each other.

The exclusion principle applies to electrons, protons and neutrons (fermions) but not photons (bosons). This can be explained in terms of spin; particles with spin $\frac{1}{2}$ obey the principle but particles with spin 1 do not.

Mass/energy equivalence

Einstein's famous equation connected energy and mass. In particle physics, the mass of a particle is often given in MeVc^{-2}. In Chapter 7 we found that a unified mass unit of 1 U is equivalent to 931.5 MeV.

Mass

Particle	Mass (MeVc^{-2})
Electron	0.5
Proton	938.3
Neutron	939.6
Neutrino	≈ 0

Table 3 Einstein established a relationship between energy and mass in his equation $E = mc^2$.

Worked example

An electron has a momentum between 1×10^{-24} Ns and 1.01×10^{-24} Ns. What is the uncertainty in its position?

Solution

The uncertainty in momentum is 0.01×10^{-24} Ns.

Applying Heisenberg's uncertainty principle $\left(\Delta p \Delta x \geqslant \frac{h}{4\pi}\right)$, the uncertainty in position is:

$$\Delta x \geqslant \frac{h/4\pi}{\Delta p} \geqslant \frac{5.28 \times 10^{-35}}{1.0 \times 10^{-26}} = 5.28 \times 10^{-9}\,\text{m}$$

Exercises

1 If you know that an electron is somewhere near a nucleus, then its uncertainty in position would be about the size of the nucleus (10^{-15} m). Calculate the uncertainty in momentum of this particle.

2 A proton has a mass of 1.0078 U. If it were changed into energy, how much energy in eV would be produced?

15.2 Fundamental interactions

Assessment statements

J.1.7 List the fundamental interactions

J.1.8 Describe the fundamental interactions in terms of exchange particles.

J.1.9 Discuss the uncertainty principle for time and energy in the context of particle creation.

Interactions between particles

Two balls travelling towards the same point will bounce off each other; the bounce is called an *interaction*. Subatomic particles interact in four different ways.

Interaction	Relative size of forces
Strong	1
Electromagnetic	10^{-2}
Weak	10^{-6}
Gravity	10^{-39}

Table 4 The four different types of interaction of sub-atomic particles.

Strong

This is the interaction between the particles of the nucleus and also their antiparticles. This force is very strong and very short range.

Electromagnetic

The electromagnetic interaction takes place between two charged particles, e.g. two electrons. This interaction becomes stronger when the distances are small, but extends to infinity $\left(F \propto \frac{1}{r^2}\right)$.

Hadrons and leptons

Hadron	Lepton
proton	electron
neutron	neutrino

Weak

Neutrinos have no charge so do not take part in electromagnetic interactions. They are also not affected by the strong nuclear force. They do not interact very often but when they do, the interaction is termed a weak interaction.

Gravity

Gravity is a force that is only significant between large masses; it is generally ignored in particle physics.

Classification of particles

The fact that different particles take part in different interactions leads to a classification of the particles: hadrons take part in strong interactions but leptons don't.

Exchange forces

In 1933 Hideki Yukawa developed the theory of exchange forces. The idea is that every force is due to the exchange of a particle. Consider two canoes paddling parallel to each other. What happens if a heavy object is thrown from one canoe to the other as shown in Figure 15.1? When A throws the ball, A must exert a force on the ball, according to Newton's third law. A will experience an equal and opposite force causing A to move away from B. When B catches the ball, B must exert a force on the ball in order to stop it. B will therefore experience a force away from A. The canoes repel each other.

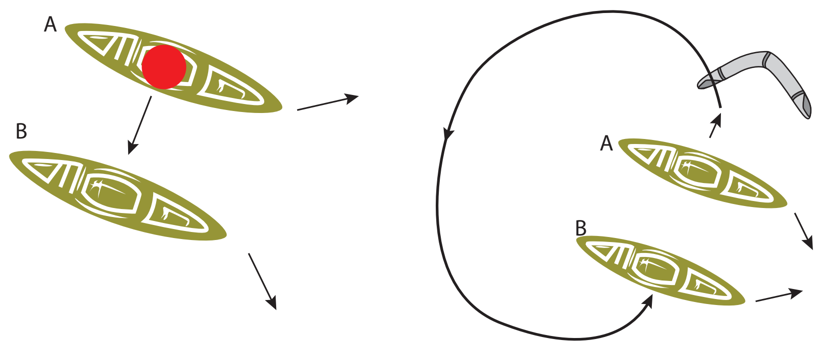


Figure 15.1 The exchange of particles leads to repulsive or attractive forces.

To explain an attractive force, imagine what would happen if A threw a boomerang in the opposite direction to B. The boomerang spins, following a path as shown in Figure 15.1. When B catches the boomerang, B will be pushed towards A.

This is not the way it happens with particles but it gives an idea that it is possible. The electric force, for example, is caused by a continual exchange of virtual photons; these interact with the charged particles to cause them to attract or repel.

Exchange particles

Yukawa linked the electromagnetic force with the photon. He explained the reason why the electromagnetic force decreases with distance $\left(F \propto \frac{1}{r^2}\right)$ in the following way: the size of the force is dependent on the number of photons that are exchanged between two particles; as photons move away from a particle they form a sphere. As the sphere gets bigger so the photons become more spread out and the force becomes weaker; since photons live for ever, they continue spreading for an infinite distance. This is why the electromagnetic force has an infinite range.

A short range force can be explained if the exchange particles disappear before they reach their objective, in other words they have a very short life. But if the particles keep disappearing, where does their energy go? Heisenberg's uncertainty principle gives the answer. Since $\Delta E \Delta t = \frac{h}{4\pi}$ if the exchange particles exist for a very short time they can have energy that doesn't have to be accounted for. Such short-lived particles are called *virtual particles*.

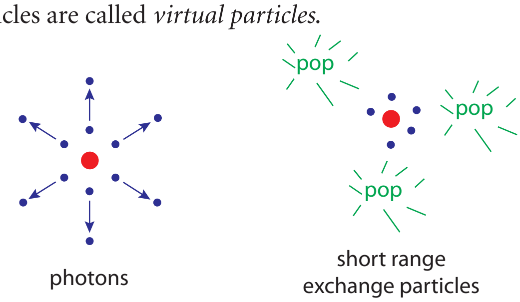


Figure 15.2 As photons spread out they cover a larger area, whereas 'short range' exchange particles disappear after a short distance.

From this it is possible to calculate the energy of the exchange particle. The force only extends within the nucleus, so the exchange particles only live long enough to travel across a small nucleus, about 1.0×10^{-15} m.

The time taken if they are travelling near to the speed of light will be

$\frac{1.0 \times 10^{-15}}{3 \times 10^{8}}\,\text{s} = 3 \times 10^{-24}\,\text{s}$

Rearranging Heisenberg's equation gives $\Delta E = \frac{h}{4\pi\Delta t}$

So $\Delta E = \frac{5.28 \times 10^{-35}}{3 \times 10^{-24}} = 1.76 \times 10^{-11}\,\text{J}$

To convert this to eV divide by 1.6×10^{-19} C

$$\Delta E = 110\,\text{MeV}$$

According to Einstein, energy and mass are equivalent, which leads us to believe that the exchange particle has a rest mass of about 110 MeVc^{-2}. This particle is called a *pion*.

Using the same argument, if the electromagnetic force is infinite, the exchange particles must live forever.

According to Einstein $\Delta E = \Delta mc^2 = \frac{h}{4\pi\Delta t}$

So if Δt is infinite, $mc^2 = \frac{h}{4\pi\infty}$ which implies that the rest mass of a photon is zero.

Range
The range of the interaction is inversely proportional to the mass of the exchange particle:

$$R = \frac{h}{4\pi mc}$$

Interaction	Exchange particle	Range	Exchange particle mass
Strong force	pion	10^{-15} m	139.6 MeVc^{-2}
Electromagnetic	photon	infinite	0
Weak	W and Z	10^{-18} m	80 GeVc^{-2}

Table 5 Summary of properties of exchange particles.

The pion takes part in strong reactions but it has a much smaller mass than the proton and neutron. This leads to a further classification as shown in Table 6.

Hadrons	
Baryons	Mesons
proton	pion
neutron	

Table 6 The pion is a meson (medium), the proton and neutron are baryons (heavy) but both are hadrons (strong).

Exercise

3 Classify the particles in Table 7 as lepton, baryon or meson.

Particle	Interaction	Mass
Muon	Electromagnetic, weak, gravity	106 MeVc^{-2}
Kaon	Strong, electromagnetic, weak, gravity	494 MeVc^{-2}
Lambda	Strong, electromagnetic, weak, gravity	1116 MeVc^{-2}
Tau	electromagnetic, weak, gravity	1777 MeVc^{-2}

Table 7

Quantum numbers and conservation

In Chapter 7 we looked at some nuclear processes. When deciding whether a certain process can take place or not we made sure that the nucleon and proton numbers were the same on each side of the equation. This is an example of a conservation principle, for example:

$$^{226}_{88}\text{Ra} \longrightarrow {}^{222}_{86}\text{Rn} + {}^{4}_{2}\text{He}$$

The proton number on the left is +88 and on the right is 86 + 2 = +88
The nucleon number on the left is 226 and on the right is 222 + 4 = 226.
Note: Nucleon number relates to the mass of the nucleus and proton number relates to its charge.

The same idea of conserving charge is used in particle physics. However, we find that charge and mass/energy is not enough, for example, it seems possible that:

$$\text{proton} \longrightarrow \text{positron} + \text{gamma}$$

In this reaction, charge and energy/mass are conserved (if the gamma has enough energy) but it never happens. A baryon can't change into a lepton. After observing all the possible interactions, a set of numbers was allocated to each particle. These numbers, so-called quantum numbers, must be conserved in all interactions.

Table 8 The quantum numbers for baryons and leptons.

	Baryon number	Lepton number
Leptons	0	+1
Antileptons	0	−1
Baryons	+1	0
Antibaryons	−1	0

The allocation of numbers is quite straight forward and illustrated in Table 8. If we apply this principle to all the particles we have mentioned so far we get the data in shown Table 9.

Table 9 This can be used to predict possible interactions.

Particle	Symbol	Charge	Baryon number	Lepton number
Photon	γ	0	0	0
Electron	e^-	−1	0	+1
Positron	e^+	+1	0	−1
Neutrino	ν	0	0	+1
Antineutrino	$\bar{\nu}$	0	0	−1
Proton	p	+1	1	0
Antiproton	$\bar{p}$	−1	−1	0
Neutron	n	0	1	0
Antineutron	$\bar{n}$	0	−1	0
Pion	π^+	+1	0	0
Antipion	π^-	−1	0	0
Pi zero	π^0	0	0	0

Examples

1 $p \rightarrow n + e^+ + \nu$

	p	$\rightarrow$	n	$+$	e^+	$+$	ν
Baryon number	1	=	1	+	0	+	0
Lepton number	0	=	0	+	$^-1$	+	1
Charge	1	=	0	+	1	+	0

Since all are conserved, this interaction seems to be possible.

2 $n \rightarrow p + e^- + \bar{\nu}$

	n	$\rightarrow$	p	$+$	e^-	$+$	$\bar{\nu}$
Baryon number	1	=	1	+	0	+	0
Lepton number	0	=	0	+	1	+	$^-1$
Charge	0	=	1	+	$^-1$	+	0

Again, all quantum numbers are conserved so the interaction is feasible.

3 $n + p \rightarrow e^+ + \bar{\nu}$

	n	$+$	p	$\rightarrow$	e^+	$+$	$\bar{\nu}$
Baryon number	1	+	1	$\neq$	0	+	0
Lepton number	0	+	0	=	1	+	$^-1$
Charge	0	+	1	=	1	+	0

Although charge and lepton number are conserved, the baryon number isn't. Therefore, this interaction is not possible.

Exercises

Use conservation principles to find out if the following are possible:

4 $p + e^- \rightarrow n + \nu$

5 $p + p \rightarrow p + p + \bar{p}$

6 $p + p \rightarrow p + p + \pi^0$

7 $p + \bar{p} \rightarrow \pi^0 + \pi^0$

8 $e^- + e^+ \rightarrow \gamma + \gamma$

9 $e^- + e^+ \rightarrow n + \gamma$

10 $p + \bar{p} \rightarrow n + \bar{\nu}$

Summary

The exchange particles we have discussed so far can be split into the following groups:

Leptons
electron
neutrino

Hadrons	
Baryons	**Mesons**
proton	pion
neutron	

Gauge bosons
photon

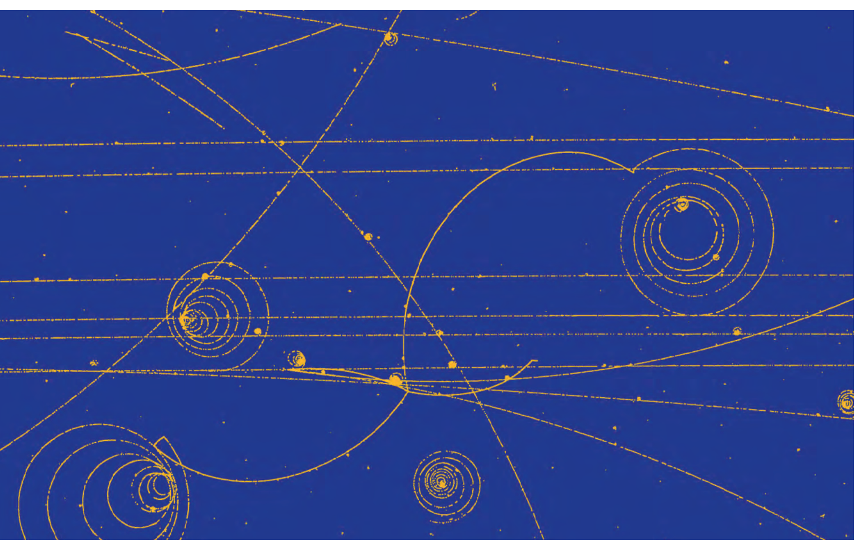
Note: Gauge bosons are particles that can occupy the same energy state; they do not obey the Pauli exclusion principle. These are the exchange particles responsible for the fundamental forces.

15.3 Particle accelerators

Assessment statements

J.2.1 Explain the need for high energies in order to produce particles of large mass.
J.2.2 Explain the need for high energies in order to resolve particles of small size.
J.2.3 Outline the structure and operation of a linear accelerator and of a cyclotron.
J.2.4 Outline the structure and explain the operation of a synchrotron.
J.2.5 State what is meant by bremsstrahlung (braking) radiation.
J.2.6 Compare the advantages and disadvantages of linear accelerators, cyclotrons and synchrotrons.
J.2.7 Solve problems related to the production of particles in accelerators.

Just as things were getting simple



A bubble chamber photograph from the CERN accelerator. There is a magnetic field directed out of the chamber causing +ve particles to spiral clockwise.

In the 1930s the picture was looking very good, as all matter could be broken down into three particles: proton, neutron and electron. The forces between these particles had been explained in terms of the exchange of pions and photons and occasionally there was an interaction that involved a neutrino. Physicists are always looking for the simplest model and this one was quite simple. However, over the next 40 years things were to change.

Cosmic rays

The best way to find out if there are any other particles in the atomic nucleus is to smash it to bits and observe the result. To do this you either have to knock the nucleus apart with high energy particles, or find a place where this is happening naturally and catch what comes out. The Sun is a place where particles are moving so fast that when they collide they can knock each other apart, and the products fly off into the universe, some even landing on Earth. A very few landed in the detectors of experimental physicists in the 1930s and 40s, enough to lead to the discovery of a whole bunch of new particles.

Particle accelerators

A particle accelerator uses an electric field to accelerate charged particles. The earliest version was simply composed of a charged sphere and an evacuated tube for the particles to travel in. The problem with early versions was that it was impossible to create a potential higher than about 7 million volts without producing a huge electrical discharge (lightning). Using a p.d. of 7 MV it is possible to accelerate particles to a KE of 7 Mev.

To overcome this barrier, particles were accelerated in stages (either around a circle as in the cyclotron and synchrotron, or in a straight line as in the linear accelerator). The principle is a bit like surfing; even on a 2 m high wave a surfer can travel downwards for several hundred metres on a long ride. By the 1950s KEs in the order of several GeV had been achieved and by 2008, the Large Hadron Collider (LHC) at CERN is predicted to produce 7000 GeV collisions.

Why high energies?

From our study of nuclear physics, we have come across the equation $E = mc^2$ which relates energy and mass. Using this equation, we can calculate the BE of a nucleus from the difference between its mass and the mass of its constituents. This will also tell us how much energy we need to put in to split a nucleus into its parts. For example, to split an iron nucleus into 26 protons and 28 neutrons requires 471.5 MeV of energy. We can also calculate how much energy we would get if we converted the mass of a particle completely into energy. An electron, for example, would be equivalent to 0.5 MeV and a proton would be 938 MeV, so it would seem possible that if we had enough energy we could create one of these particles.

Creating particles from energy

A photon of light has an energy given by the equation $E = hf$ but it has no mass. Using this equation, we can calculate that a photon of frequency 1.2×10^{20} Hz has an energy of 0.5 MeV. It would therefore seem possible to turn this energy into an electron. However, if this happened we would have violated not only the law of conservation of charge but also lepton number; this cannot therefore be possible. What is possible is to convert a 1.0 MeV photon into an electron-positron pair; this is called *pair production*. If this happens then the energy of the photon is converted into two particles that have mass but no KE. If they are not moving, then they have no momentum, so what has happened to the momentum of the photon? The only way that this can happen is if it happens next to another particle, for example, a nucleus that can gain the momentum lost by the photon.

Momentum of a photon
Even though a photon has no mass, it has momentum. This is a consequence of the theory of relativity dealt with in Chapter 13.

The magnitude of the momentum, p is given by the equation $p = \frac{h}{\lambda}$.

So if a positron and an electron can be created from a high-energy photon, maybe other heavier unknown particles could be created from higher energy photons. Such high-energy photons can be created by doing the reverse of the above. If a positron collides with an electron then photons are produced. It's not only the mass that is converted to photon energy but also any KE the particles had before colliding – so to produce high-energy photons you need high-energy electrons and positrons.

Resolution

In Chapter 4 you learnt that when light passes through a small aperture it diffracts. This spreading out of light can make it difficult to resolve two close points but it can also be useful for finding information about the aperture. However, to find information, the wavelength must be about the same size as the aperture. If we take the example of a narrow slit of width d, we know that the angle θ at which the first minimum occurs is given by the equation $d \sin \theta = \lambda$. If we can measure θ then we can calculate d. However, if d is very small, the light is diffracted into

When electrons are accelerated to such high speeds, we need to apply relativity to calculate their momentum and KE. The equation relating momentum and energy is

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

However, for high-energy electrons this approximates to $E = pc$.

a circular wavefront, so we can't measure θ and can therefore not obtain any information about d except that it is small. So if we are to use diffraction to find out information about the particles in the nucleus, we need to use waves that have a wavelength about the size of those particles. The nucleus has a diameter in the order of 10^{-15} m, so to find useful information about the nucleus we would need to use photons of similar sized wavelength. A photon of such short wavelength would have an energy of 1 GeV. Alternatively high-energy electrons could be used. An electron with momentum p has a de Broglie wavelength equal to h/p, and to create an electron of such short wavelength it would have to be accelerated to a KE of about 1 GeV.

Exercises

11 Estimate the minimum energy required to produce a proton–antiproton pair.

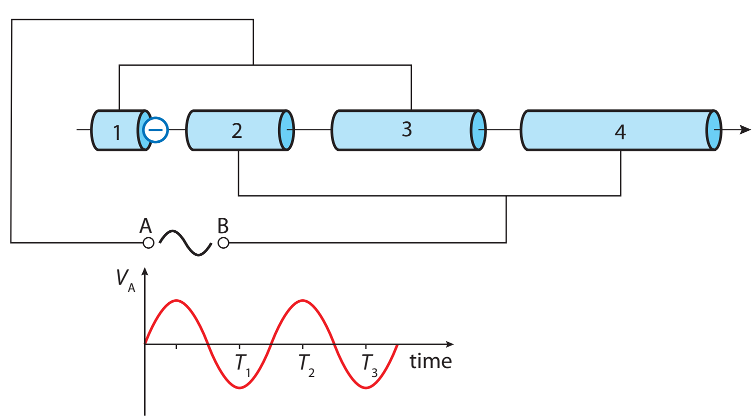
12 An electron is accelerated by a p.d. of 1000 V. Calculate
- **(a)** its KE in eV
- **(b)** its KE in joules
- **(c)** its momentum (use classical mechanics)
- **(d)** its de Broglie wavelength.

13 An electron is accelerated to an energy of 20 GeV. Calculate
- **(a)** its momentum (use the relativistic equation in the box)
- **(b)** its de Broglie wavelength.

The linear accelerator (linac)

In a linear accelerator, bunches of charged particles are accelerated as they travel along a line of straight tubes. The actual acceleration takes place as they pass from one tube to the next, whilst in the tubes the particles have constant velocity. The trick is to make sure that the tubes have the right polarity at the right time or the particles would be slowed down instead of accelerated. Figure 15.3 shows how this is achieved.

Figure 15.3 A linear accelerator with 4 tubes and 3 acceleration gaps.



The p.d. supplied to the tubes is alternating, so tube 1 will sometimes be positive and sometimes negative. The graph shows the potential of terminal A with time; this is the same as tubes 1 and 3 since they are connected to this terminal. Electrons enter tube 1 along the axis of the tube. When they get to the gap between tubes 1 and 2, tube 1 must be negative . So if the electrons arrive at time T_1 they will be accelerated. As the electrons enter tube 2 they are not affected by the field, so continue with constant velocity to the gap between tube 2 and 3. Now they will be accelerated if tube 3 is positive, which is at time T_2 , then when they come out of tube 3 it must be negative again, so the time must be T_3. So the timing must be just right to make sure that each time the electrons are in the gap they are accelerated. This is achieved by making the tubes the right length so that the time taken for the p.d. to change direction is the same as the time for the electrons to travel through the tube. In this case the time should be $T_2 - T_3$ (half the time period). An alternative would be to increase the frequency of the signal as the electrons get faster. So if the AC frequency is 1 MHz, the particles will spend 5×10^{-7} s in each tube. If the p.d. across each gap is 500 kV then, with three gaps, the electrons will gain 1.5 MeV of energy. To increase this energy would require the addition of more sections, each section getting longer and longer. The Stanford Linear Accelerator is 3 kilometres long.

The cyclotron

One problem with a linac is that to produce high-energy particles it has to be very long; a cyclotron gets around that problem by using a magnetic field to make the particles move in a circle. Instead of straight tubes, the particles travel in two D shaped vacuum-filled chambers and are accelerated as they pass from one to the other. Figure 15.4 shows how these are arranged. The chambers are placed in a uniform magnetic field, causing charged particles introduced near the centre to travel in a circular path. Each time the particles pass across the gap they are accelerated, and as this happens the radius of their path increases, resulting in a spiral path as shown in Figure 15.5.

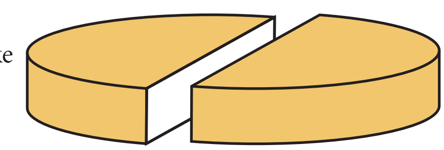


Figure 15.4 The hollow D shaped containers in which the charges travel.

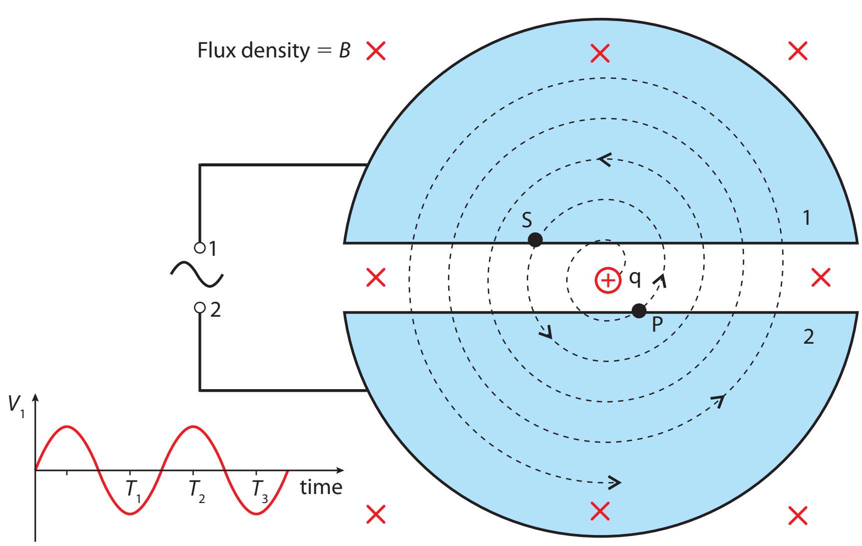


Figure 15.5 The path of a positive particle as it is accelerated around a cyclotron. The graph shows the changing potential of D_1.

Starting from the centre we can see by applying Flemings' left hand rule, that the positive particle experiences a magnetic force causing it to move in the circular path shown. If the particle gets to point P at time T_1 then D_1 will be negative so it will be accelerated, resulting in a circular path of bigger radius. If at time T_2 the particle is at S then D_1 will be positive so the particle is again accelerated. So provided that the time between each crossing of the gap is half the time period of the AC source then the particle will accelerate each time. This is just the same as with the linac except, instead of having longer and longer tubes, the path length is automatically adjusted as the circles get bigger. We can see how that works out by looking at the equations of the motion.

Whilst in the D chambers, the motion is circular, the centripetal force being provided by the magnetic force.

$$\frac{mv^2}{r} = Bqv$$

So the radius of the path is

$$r = \frac{mv}{Bq}$$

Now the speed, $v = \frac{2\pi r}{T}$ where T is the time period. Substituting for v gives:

$$r = \frac{m2\pi r}{BqT}$$

Rearranging gives T, the cyclotron period.

$$T = \frac{2\pi m}{Bq}$$

So the time period is independent of the radius. As the radius gets bigger the particle takes exactly the same time to go round the D.

Rearranging the first equation gives

$$v = \frac{Bqr}{m}$$

So the KE is given by:

$$\tfrac{1}{2}mv^2 = \frac{B^2q^2r^2}{2m}$$

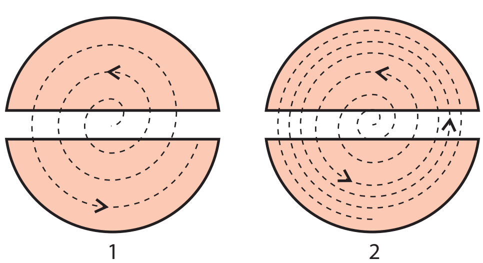


Figure 15.6 Two cyclotrons with different accelerating p.d.s.

The maximum KE will be when the particle has a radius equal to the D radius. In Figure 15.6, cyclotron 2 has a lower accelerating p.d. than cyclotron 1, and each time the charges cross the gap in 1 they gain more KE. However, this causes the spiral to unwind more quickly resulting in less gap crossings. The result is that the particles in each cyclotron have the same final KE.

The cyclotron is widely used for producing particles with energies in the MeV region. However, to produce higher energies would need larger D chambers and very big magnets. These are difficult to produce, remember the chambers are vacuum-filled and the magnetic field covers the whole area.

Synchrotron

The operation of the synchrotron is similar to the linac except that the tubes are arranged in a circle as in Figure 15.7; this means that the particles can be sent around more than once, increasing the amount of energy gained without increasing the length of the accelerator.

In Figure 15.7, eight very strong superconducting magnets are used to bend the beam in a circle. Between these magnets the particles are accelerated as they pass through an electric field. As the particles get faster and faster the time spent in each section becomes smaller and smaller, so to compensate for this, the frequency of the accelerating potentials must be increased. Also, as they get faster the force needed to hold the particles in a circle increases. To correct this, the current through the magnets is increased. It is important that the increase in current is synchronised with the increasing frequency, hence the name synchrotron.

Superconducting magnets
Increasing current in an electromagnet increases the flux density; however it also produces heat, which limits the maximum flux. At very low temperatures some materials become superconducting; this means they have no resistance so produce no heat when current flows. So superconducting electromagnets can produce very high fields without getting hot.

Increasing the energy of the particles is simply a matter of sending them around the ring more times. As the particles approach the speed of light, their speed does not increase significantly, only their mass. So the limiting factor is the flux density of the B field. This factor limits the maximum energy attainable by a given synchrotron; to get higher energies would require a bigger circle. The biggest synchrotron is at the European Centre for Nuclear Research (CERN) in Geneva, with a circumference of 27 kilometres.

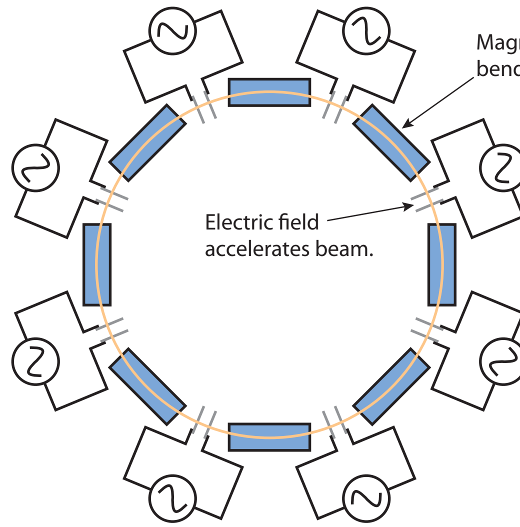


Figure 15.7 The synchrotron.

To do research into particle physics requires high energies, and high energies mean big accelerators, which cost a lot of money. The latest addition at CERN (the LHC) is said to cost in the region of 10 billion US dollars. This means that either only rich countries can do this research, or countries must share resources and results. The facility at CERN is a cooperation between more than 80 countries.

For more information on CERN, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 15.1.

Synchrotron radiation (Bremsstrahlung)

The charged particles travelling around the synchrotron accelerate each time the direction of their path is changed by the magnetic field. When a charged particle accelerates it emits electromagnetic radiation so EM radiation in the X-ray region of the spectrum is emitted on each bend of the synchrotron, resulting in a loss of energy. This X-radiation is very intense, polarized and parallel which makes it particularly suitable for use in X-ray diffraction and other experiments investigating the properties of materials. Some synchrotrons are in fact specifically constructed as a source of this radiation.

Collision experiments

One advantage that the synchrotron has over other types of experiments is that the same ring can be used to accelerate particles of opposite charge (protons and antiprotons). Once they reach sufficient energy, they can be collided with each other. A collision between two beams of fast moving targets will release much more energy than the beam hitting a fixed target.

Advantages and disadvantages
Since the Linac beam is straight there is no radiation loss on the bends as there is in a synchrotron. However, acceleration in a linac is a one off event, if collisions don't take place at the end of the track then the particles are lost, particles in a synchrotron go round and round so there are more chances.
Due to difficulties constructing big enough Ds and magnets the cyclotron is not a serious contender for very high energies; however it is often used as a source of fast moving particles in other applications.

Energy released

Head-on collision If two particles of equal mass, e.g. a proton and antiproton, collide head-on, the energy available for the production of new particles = rest energy + KE of both particles.

Stationary target If one particle collides with a stationary target, not all the energy can be converted to new particles. This is because to conserve momentum, the new particles must move away in the same direction as the projectile. The energy available is then given by the formula:

$$E_A^2 = 2EMc^2 + (Mc^2)^2 + (mc^2)^2$$

where E_A = the energy available
M = rest mass of the target
m = the rest mass of the accelerated particle
E = the total energy of the accelerated particle
All energies are measured in eV.

Worked example

How much energy would be available if a 10 GeV proton collided with a stationary proton?

Solution

The calculation would be rather awkward if we were to use kg, m and s, but if we use eV throughout, it is fairly straightforward.

Proton rest mass $= 938\,\text{MeV}c^{-2} = 0.938\,\text{GeV}c^{-2}$

so $E_A{}^2 = 2 \times 10 \times 0.938 + 0.938^2 + 0.938^2$

$E_A = 4.5\,\text{GeV}$

Exercises

14 To create a particle of total rest mass 3 GeV c^{-2} from the collision of two protons, how much energy would the accelerated proton have to be given?

15 After a high-energy collision, two protons can form a pion. The equation for this process is:

$p + p \rightarrow p + p + \pi$

If the rest mass of the pion is 135 MeV, calculate

(a) how much energy must be available for this to take place
(b) the total energy of the accelerated proton
(c) the KE of the proton (this is total energy – rest energy).

16 The following equation shows how two high energy protons can form an antiproton assuming one of the protons is stationary.

$p + p \rightarrow p + p + p + \overline{p}$

(a) Calculate
  (i) the energy that must be available for the formation of the products
  (ii) the total energy of the accelerated proton
  (iii) the KE of the proton.
(b) If the particles were moving towards each other, calculate the KE of each.

17 In the Large Hadron Collider at CERN the beam consists of 3000 bunches of particles accelerating around the ring at the same time. Each bunch contains 10^{11} particles and each particle has an energy of 7 TeV.

(a) Calculate the total energy of the beam in joules.
(b) If a 50 000 tonne ship had a KE the same as the total beam energy, how fast would it be travelling?

15.4 Particle detectors

Assessment statements

J.2.8 Outline the structure and operation of the bubble chamber, the photomultiplier and the wire chamber.
J.2.9 Outline international aspects of research into high-energy particle physics.
J.2.10 Discuss the economic and ethical implications of high-energy particle physics research.

Providing enough energy to create new particles is only the first step; the next is to identify those particles and confirm that they are new. In Chapter 7 you read about the cloud chamber, which is used to enable us to see the paths of particles as they pass through a vapour. This has limited use due to the fact that it is filled with a gas that moves around, making it difficult to say whether a change in direction of a track is due to a change in the particle or a movement of the gas. There are also many particles that will simply pass through the gas without detection. For these reasons, liquid filled or solid state detectors are often used when observing the products of high-energy collisions; these detectors are arranged around the site of a collision ready to detect anything that flies out.

High energy physics is like trying to find out what is in a box by putting it on a hill several kilometres away, firing a rocket at it and standing with your hands open waiting to catch something. But you have to remember that if you blew up millions of boxes you might eventually catch something.

The bubble chamber

If you have ever watched bubbles rising in a glass of champagne or lemonade you may have noticed that the bubbles always rise from certain points at the bottom or sides of the glass. These places are where there are small bits of dust stuck to the glass; the bubbles form around the impurities and rise to the top. In a bubble chamber, bubbles form around the ions created along the track of a high-energy particle, enabling the path of the particle to be observed. The liquid used in the bubble chamber is hydrogen that has been cooled below its boiling point. The reason that champagne has bubbles is because of the dissolved carbon dioxide it contains, but to make hydrogen liquid form bubbles, it needs to be heated until it boils. Alternatively, the pressure can be reduced, lowering the boiling point, and this is what happens in the bubble chamber as shown in Figure 15.9.

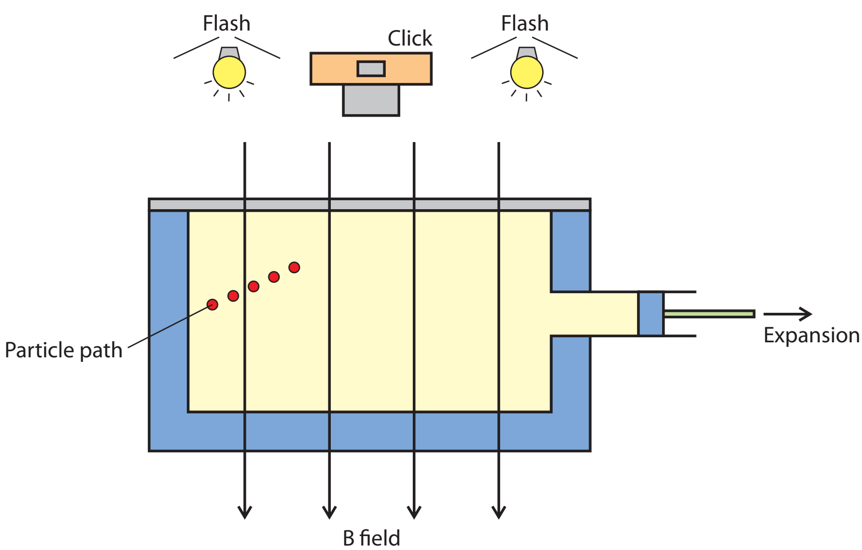


Figure 15.9 A bubble chamber.

Liquid hydrogen is contained in a pressurized container at a temperature below its boiling point. The pressure is then suddenly released, lowering the boiling point; the hydrogen should now be gas. However, it can't suddenly change to gas but forms bubbles around any impurities, such as ions formed by a passing high-energy particle. As this happens a photograph is taken through the glass top. The whole chamber is placed in a magnetic field, which causes charged particles to be deflected; this helps in identifying the particles' charge and momentum.

There are a couple of problems with the bubble chamber that make it unsuitable for very high-energy particles. High-energy particles often interact to produce particles that fly off in all directions at very high speeds. These particles might then have further interactions over a large area, and to catch all these interactions the detector must be very big, but it is not possible to make very large bubble chambers. Another problem is that the expansion of the hydrogen must be timed to coincide with the entry of a particle; once triggered the chamber must rest before it is used again. Finally in this digital age it would be much more convenient to have a digital readout rather than a photograph.

A bubble chamber photograph showing how the tracks are curved by the magnetic field.

The wire chamber

When ions are formed in a gas they enable it to conduct electricity. This property is used in many particle detectors, including the Geiger-Muller tube and wire chamber. The GM tube detects when a particle passes through it but not which way it went. If you have a room full of GM tubes you could track the path too, and that's basically what a wire chamber is.

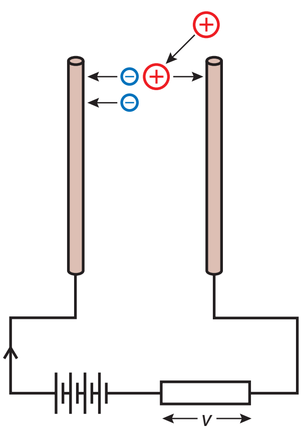
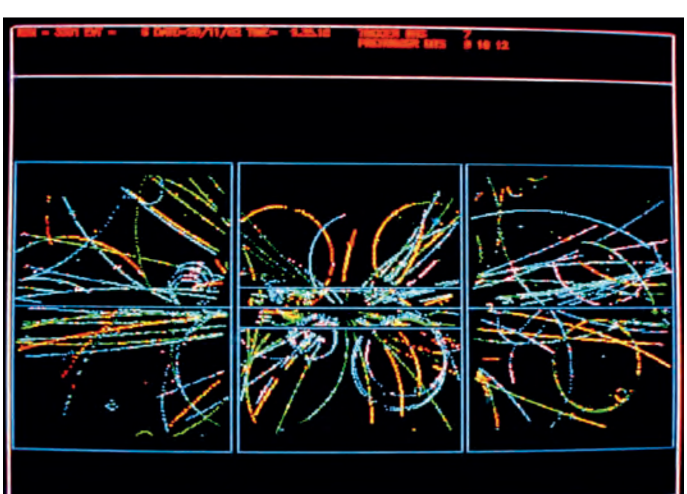


Figure 15.9 A gas atom is ionized by a high energy particle.

Figure 15.9 shows the principle behind the operation of the wire chamber. A battery creates an electric field between two wires, when a gas atom is ionised between the wires, the electrons flow towards the positive potential. This causes a current to flow in the circuit, leading to a rise in p.d. across the resistance. It is this rise in p.d. that is the signal that a particle has passed this way.

Multiwire chamber

In a multiwire chamber, an array of wires can be used to determine the position of a particle. The output from each wire is easily converted to a digital signal, which can be analysed by a computer. To gain a 3D picture of particle motion, identical grids are arranged at 90° to each other; the computer can then calculate the coordinates of a passing particle and produce a 3D representation of the path.



A 3D computer generated picture from a wire chamber at CERN. The different colours give the depth of the tracks – red and yellow are near, blue and green are more distant.

Drift chamber

An alternative way of tracking a particle in 3D is to arrange the wires much further apart and use the time for the electrons to reach the wires to calculate the distance the particle is from the wire.

The parallel wires in a drift chamber.

Photomultiplier

Some of the products of high-energy collisions are gamma rays and, although a beam of gamma rays is very ionizing, a single one isn't. Remember that if a photon interacts with matter it disappears completely, so a gamma ray photon does not leave a track of ions as an electron would. To detect single photons, a photomultiplier can be used as in Figure 15.10.

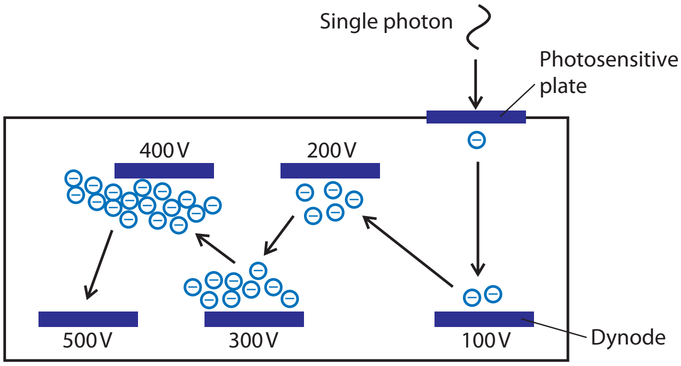


Figure 15.10 A photomultiplier.

A single photon enters the photomultiplier through a small window coated on the inside with a material that absorbs the photon and emits an electron (the photoelectric effect). The electron is then accelerated towards the first of several positive dynodes. Each time the electrons collide with a dynode, more electrons are liberated, resulting in an amplification of the signal. By the end of the tube the signal is big enough to cause a measurable change in potential.

Scintillator

High-energy gamma rays will not result in the emission of electrons from a photosensitive surface so cannot be detected directly with a photomultiplier. However, there are some materials that will emit visible light when they absorb gamma radiation; this is called scintillation. If a piece of this material is placed before the window then detection is achieved.

Exercises

18 How many joules of energy is 7000 GeV equivalent to?

19 If a bunch of particles contains 10^{12} protons, each with an energy of 7000 GeV, how much energy would the beam have?

20 If a 1 kg mass had the same amount of kinetic energy, how fast would it be moving?

21 If a bunch of particles hits a target every 10^{-29} seconds, what is the power of the beam?

22 Figure 15.11 shows the tracks of an electron and a positron produced from a high energy gamma photon. (These are the spiralling tracks A and B, the other is an electron that has been knocked out of an atom of hydrogen.) The tracks are curved because there is a magnetic field into the plane of the page. From your knowledge of magnetic fields, deduce which particle is the electron and which is the position.

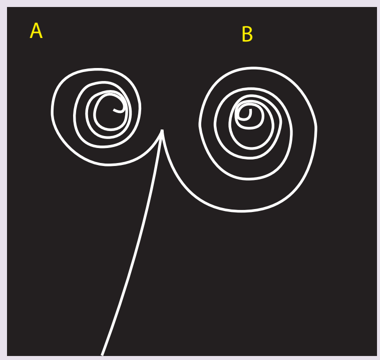


Figure 15.11 A bubble chamber image of electron pair positron pair production.

The particle explosion

Once physicists started using the correct tools to look in the right places, they started to find the particles they were looking for; in fact they found hundreds of them. As each particle was discovered, they were classified according to their interactions. The following tables list some of these particles. Remember, each particle has an antiparticle with opposite charge.

Table 10 Many new hadrons have been discovered and this table includes some of the mesons.

Mesons			
Name	Symbol	Baryon	Strangeness
Pion	π^+	0	0
Pion	π^o	0	0
Kaon	κ^+	0	1
Kaon	κ^o	0	1
Eta	η	0	0
Rho	ρ^+	0	0
J/psi	ψ	0	0
Phi	ϕ	0	0
D	D	0	0

Strangeness
Some particles do not interact as predicted by the conservation of charge and baryon number. This strange behaviour led to the introduction of a new quantum number '*strangeness*'.
λ and Σ have strangeness of -1
Ξ have strangeness of -2
K has strangeness of $+1$
Strangeness is conserved in strong and electromagnetic interactions but not weak interactions. Strange!

Baryons			
Name	Symbol	Baryon	Strangeness
Proton	p	1	0
Neutron	n	1	0
Delta	Δ^{++}	1	0
Delta	Δ^{+}	1	0
Delta	Δ^{o}	1	0
Delta	Δ^{-}	1	0
Lambda	λ^{o}	1	−1
Sigma	Σ^{+}	1	−1
Sigma	Σ^{o}	1	−1
Sigma	Σ^{-}	1	−1
Xi	Ξ^{o}	1	−2
Xi	Ξ^{-}	1	−2
Omega	Ω^{-}	1	−3

Table 11 Many new hadrons have been discovered and this table includes some of the baryons.

Leptons			
Charged lepton		Uncharged leptons	
Electron	e	Electron neutrino	ν_e
Muon	μ	Muon neutrino	ν_μ
Tau	τ	Tau neutrino	ν_τ

Table 12 Relatively few leptons have been discovered compared to the number of hadrons.

Lepton number
In any interaction there must be conservation of muon lepton number, electron lepton number and tau lepton number.
Leptons have lepton number +1 and anti leptons have lepton number −1.

Exercises

23 Given the information that strangeness is not conserved in weak interactions, deduce which of the following interactions are weak:

(a) $\Sigma^{+} \rightarrow p + \pi^{o}$
(b) $\Sigma^{o} \rightarrow \lambda^{o} + \gamma$
(c) $K^{o} \rightarrow \pi^{+} + \pi^{o}$

24 Is the lepton number conserved in the following interactions?

(a) $n \rightarrow p + e^{-} + \overline{\nu}$
(b) $e^{-} \rightarrow \mu + \gamma$
(c) $\mu \rightarrow e^{+} + \overline{\nu}_e + \nu_\mu$

15.5 Quarks

Assessment statements

J.3.1 List the six types of quark.
J.3.2 State the content, in terms of quarks and antiquarks, of hadrons.
J.3.3 State the quark content of the proton and the neutron.
J.3.4 Define baryon number and apply the law of conservation of baryon number.
J.3.5 Deduce the spin structure of hadrons.

Quarks

By 1960 the particle model was looking quite complex and rather untidy. It would simplify matters greatly if all the hadrons were made out of a few common particles. In 1964 Murray Gell-Mann proposed that all hadrons were made of combinations of fundamental particles called quarks.

Spin

As mentioned, particles have a property called spin. Baryons have spin $\frac{1}{2}$ or $1\frac{1}{2}$ and mesons 0 or 1. Spins can either add, if they are in the same direction, or cancel, if opposite. If quarks have spin $\frac{1}{2}$ then baryons can be made of 3 quarks and mesons can be made of 2, as shown.

$$\overset{\uparrow}{+\tfrac{1}{2}}\ \overset{\downarrow}{-\tfrac{1}{2}}\ \overset{\uparrow}{+\tfrac{1}{2}} = $$

$$\overset{\uparrow}{+\tfrac{1}{2}}\ \overset{\uparrow}{+\tfrac{1}{2}}\ \overset{\uparrow}{+\tfrac{1}{2}} = 1\tfrac{1}{2}$$

$$\overset{\uparrow}{+\tfrac{1}{2}}\ \overset{\downarrow}{-\tfrac{1}{2}} = 0$$

$$\overset{\uparrow}{+\tfrac{1}{2}}\ \overset{\uparrow}{+\tfrac{1}{2}} = 1$$

Neutron

A neutron is a baryon so contains 3 quarks. It has zero charge so the charges of the quarks must cancel e.g. $+\frac{2}{3}-\frac{1}{3}-\frac{1}{3}$

Proton

A proton also contains 3 quarks but has a charge of $+1$. Assuming these are the same types of quark that make up a neutron, we can achieve this with $+\frac{2}{3}+\frac{2}{3}-\frac{1}{3}$

So we can make the proton and the neutron from combinations of 2 different types of quark. The quarks are said to have different flavours, the $-\frac{1}{3}$ has flavour down and the $+\frac{2}{3}$ flavour up.

To make all known hadrons we need a total of 6 flavours.

Table 13 The six possible quarks that make up hadrons.
Note: To keep things simple, charm, bottom and top quarks will not be considered in the following examples.

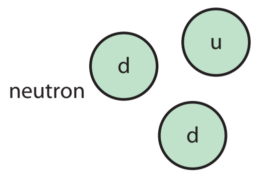
Quark			Antiquark	
Flavour	Symbol	Charge	Symbol	Charge
Up	u	$+\frac{2}{3}$	$\bar{u}$	$-\frac{2}{3}$
Down	d	$-\frac{1}{3}$	$\bar{d}$	$+\frac{1}{3}$
Strange	s	$-\frac{1}{3}$	$\bar{s}$	$+\frac{1}{3}$
Charm	c	$+\frac{2}{3}$	$\bar{c}$	$-\frac{2}{3}$
Bottom	b	$-\frac{1}{3}$	$\bar{b}$	$+\frac{1}{3}$
Top	t	$+\frac{2}{3}$	$\bar{t}$	$-\frac{2}{3}$

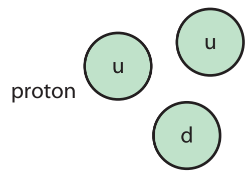
Strangeness

Strange particles contain strange quarks. The strangeness number is calculated from the number of antistrange quarks – strange quarks.

Baryons

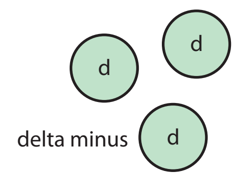
Baryons are made out of three quarks.





Delta minus

	Charge	Baryon number	Strangeness
$\Delta-$	-1	1	0

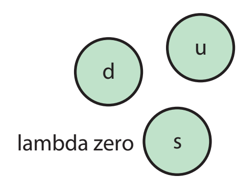


Quark combination:

	d	d	d
Charge	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$

Lambda zero

	Charge	Baryon number	Strangeness
λ^{o}	0	1	-1



Strangeness -1 implies strange quark.

The combination *uds* has the correct charge and strangeness.

Quark combination:

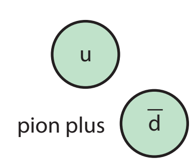
	u	d	s
Charge	$\frac{2}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$

Mesons

Mesons are a combination of a quark and an antiquark.

Pion plus

	Charge	Baryon number	Strangeness
π^{+}	$+1$	0	0

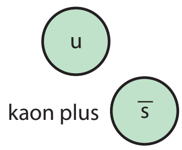


Quark combination:

Strangeness 0 implies no strange quarks.

The combination up, antidown will give a charge $+1$.

	u	$\bar{d}$
Charge	$+\frac{2}{3}$	$+\frac{1}{3}$



Kaon plus

	Charge	Baryon number	Strangeness
K^+	$+1$	0	$+1$

Quark combination:

Strangeness $+1$ implies that the K^+ has an antistrange quark.

An antistrange quark has charge $+\frac{1}{3}$ so to give a total charge of $+1$ an up quark is needed so combination is $u\bar{s}$:

	u	$\bar{s}$
Charge	$+\frac{2}{3}$	$+\frac{1}{3}$

To test your predictions, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 15.2.

Exercise

25 Using the properties of the particles found in the above tables, deduce the quark content of the following particles:

(a) π^-

(b) Ω^-

(c) Ξ^-

(d) Ξ°

26 During β-decay a neutron changes to a proton. Which quark changes during this process?

Properties of quarks

Quark confinement

No one has ever detected a free quark; they only exist confined inside other particles. This phenomenon is called *quark confinement*. The reason for this is due to the nature of the force between quarks, which is very strong and constant with increasing distance (unlike electromagnetic force that gets less). As two quarks are separated, the energy transferred gets bigger and bigger (work done = force × distance). At a certain point the energy becomes big enough to create another quark or antiquark, so instead of a free quark you get a meson.

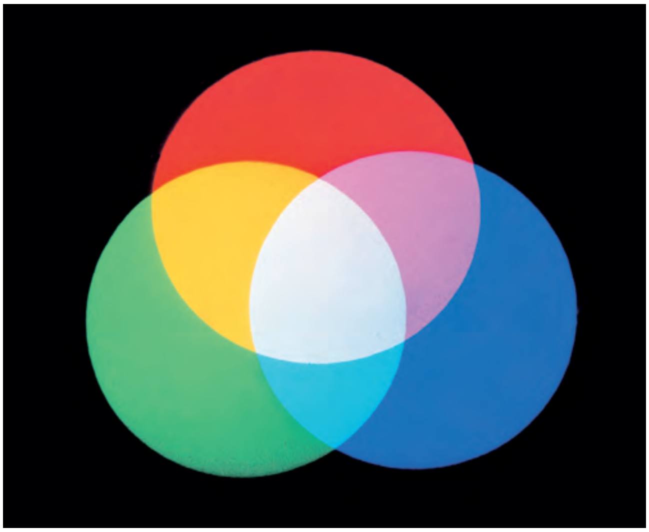
Since we can never detect a quark on its own, can we really say that they exist?

Gluon

We have seen that the forces between sub-atomic particles are exchange forces: electromagnetic force is due to the exchange of photons and the strong force is due to the exchange of pions. The force between quarks is also the strong force. However, the exchange particle in this case is the gluon rather than the pion. This is a bit confusing but we have come across a similar case when talking about the force between atoms. This is called the interatomic force, even though it is actually the electric force between the electron clouds of the different atoms. We call this a *residual force*. So the force between nucleons is due to the exchange of pions, but it is really the gluons that are exchanging between the pions and nucleons that transfer the force.

Colour charge

To explain the observation that two electrons repel each other we used the concept of charge – this is the name given to the property/thing that causes the force. The fact that some particles attract each other and some repel leads us to believe that there are two types of charge. This is very convenient because we also have two types of numbers + and –, so we use + and – to represent charge. Even more convenient is the way that opposite charges cancel in the same way that numbers do. If charge is the property that causes the electromagnetic interaction, what property of quarks causes the strong interaction? This time it isn't so simple, as there are three types of this property not two, so we can't use numbers. However, there are three primary colours on which the model can be based, and the property is called *colour charge*.



Red, green and blue light combine together to give white light.

Quarks can be red, green or blue, and antiquarks are anti-red, anti-green and anti-blue. All particles must be white, so a baryon must contain a red, green and blue and a meson, for example, red and anti-red. Gluons, however, have two colours giving them colour charge and the ability to carry the strong force. The complete theory is called *Quantum Chromo Dynamics* (QCD).

QCD is an example where numbers cannot be used to model the physical property.

The fact that quarks in a particle have different colours allows them to exist in the same quantum state. This would otherwise not be allowed due to the Pauli exclusion principle, which states that particles with spin $\frac{1}{2}$ cannot exist in the same state.

Fundamental particles

The standard model

A fundamental particle is one that cannot be broken down into anything smaller; a proton is therefore not fundamental but an electron is. The standard model of particle physics splits the fundamental particles into three groups: leptons, quarks and exchange particles (gauge bosons). Tables of these particles show a certain degree of symmetry. The particles are ordered in three groups according to mass (heaviest to the left) – these groups are called *generations*.

The standard model is very simple, you could say it is beautiful, but is the simplest model necessarily correct?
Why are we always trying to find the simplest solution?

Quarks

Charge	1st Generation	2nd Generation	3rd Generation
$+\frac{2}{3}$	up	charm	top
$-\frac{1}{3}$	down	strange	bottom

Table 14 Each of these fundamental particles has an anti-particle, which can also be arranged in a similar way.

Leptons

Charge	1st Generation	2nd Generation	3rd Generation
0	electron neutrino	muon neutrino	tau neutrino
−1	electron	muon	tau

Table 15 Each of these fundamental particles has an anti-particle, which can also be arranged in a similar way.

Gauge bosons

Table 16 These are the exchange particles that are responsible for the fundamental forces.
Note: it is now recognised that at high energies the electric and weak forces are one and the same thing. This is called the electroweak force.

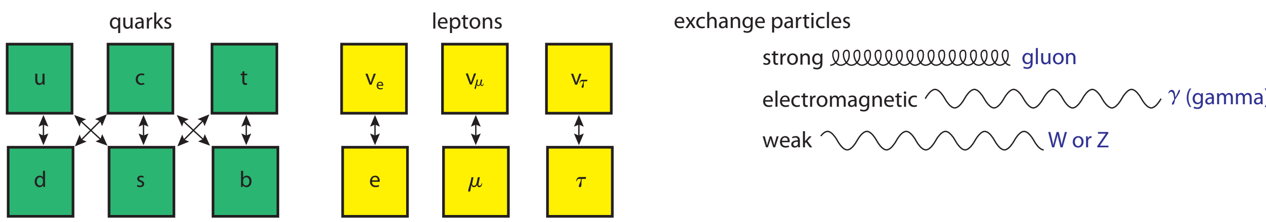
Particle	Force
Gluon	Strong
Photon	Electromagnetic
W and Z	Weak

Interactions

Whenever an interaction takes place an exchange particle is either emitted or absorbed. The type of particle absorbed determines the type of interaction (e.g. photon if electromagnetic).

Leptons can only interact with leptons in the same generation.
Quarks can only interact with quarks in the same generation or diagonal (e.g. up and strange).

Figure 15.12 The standard model of possible particle interactions.



Exercises

27 Can an electron neutrino interact with a tau?

28 Can a muon interact with a muon neutrino?

29 Can a charm quark interact with a down quark?

30 Can an up quark interact with a bottom quark?

31 Which exchange particle will be involved with an interaction between a top and a bottom quark?

32 Which exchange particle is involved in an interaction between an electron neutrino and an electron?

15.6 Feynman diagrams

Assessment statements

J.1.10 Describe what is meant by a *Feynman diagram*.

J.1.11 Discuss how a Feynman diagram may be used to calculate probabilities for fundamental processes.

J.1.12 Describe what is meant by *virtual particles*.

J.1.13 Apply the formula for the range *R* for interactions involving the exchange of a particle.

J.1.14 Describe pair annihilation and pair production through Feynman diagrams.

J.1.15 Predict particle processes using Feynman diagrams.

Feynman diagrams represent the particle interactions that are possible according to the standard model summarized in Figure 15.12.

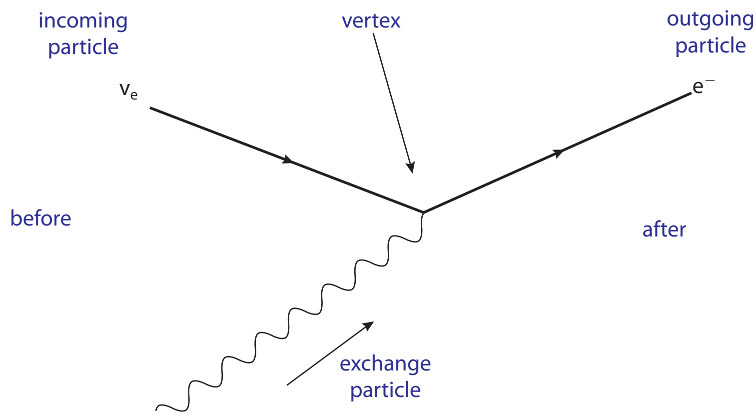


Figure 15.13 The Feynman diagram of an electron and an electron neutrino shows that a particle comes in, absorbs an exchange particle and a particle comes out. So, before the interaction there was a ν_e and an exchange particle after was an e^-.

Feynman diagram interactions

It is important to realise that a Feynman diagram is not a drawing of the particle paths, it is just a representation of the interaction.

- The vertex of the diagram is always made up of two particles and an exchange particle.
- Particles have straight lines and exchange particles have wavy lines.
- Time progresses from left to right.
- The direction of the particles is given by the arrows, antiparticles travel in the opposite direction to the arrow.

Feynman diagram rules

- Each vertex has two arrows and one wavy/curly line.
- Particle arrows point forwards in time and antiparticles backwards.
- There's always one arrow entering and one leaving a vertex.
- You can't mix leptons and quarks on one vertex.

According to the standard model, an electron should interact with an electron neutrino. The interaction will be a weak interaction since the neutrino can only take part in weak interactions (see Figure 15.15).

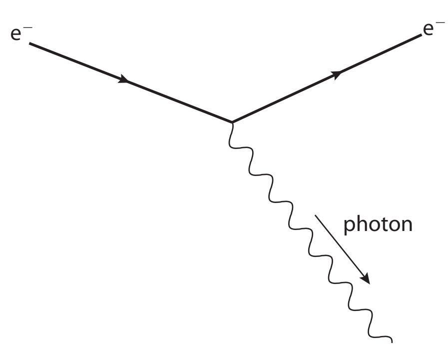


Figure 15.14 An electron emits a photon: notice the vertex has two arrows and one wavy/curly line.

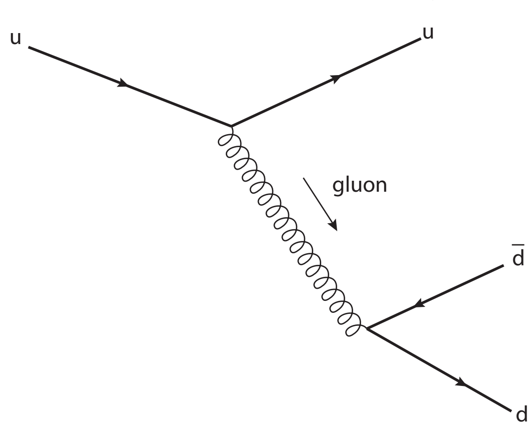


Figure 15.15 An up quark emits a gluon that then forms a quark/anti-quark pair: notice the antiquark arrow points backwards in time.

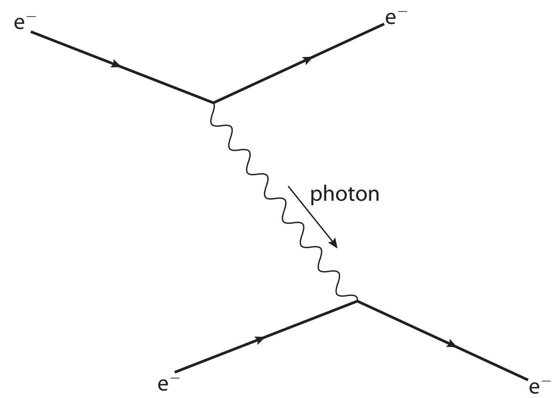


Figure 15.16 An electron emits a photon absorbed by another electron. This is the electromagnetic repulsion between two electrons.

Making predictions

Feynman diagrams are not just a nice way to represent interactions, they can also be used to make predictions. If we take a Feynman diagram and swing the arrows around it will give the diagram for another interaction – all such interactions are possible. For example, Figure 15.17 shows an electron emitting a photon. If we swing the arrows around (still following the rules) we can get another two possible interactions.

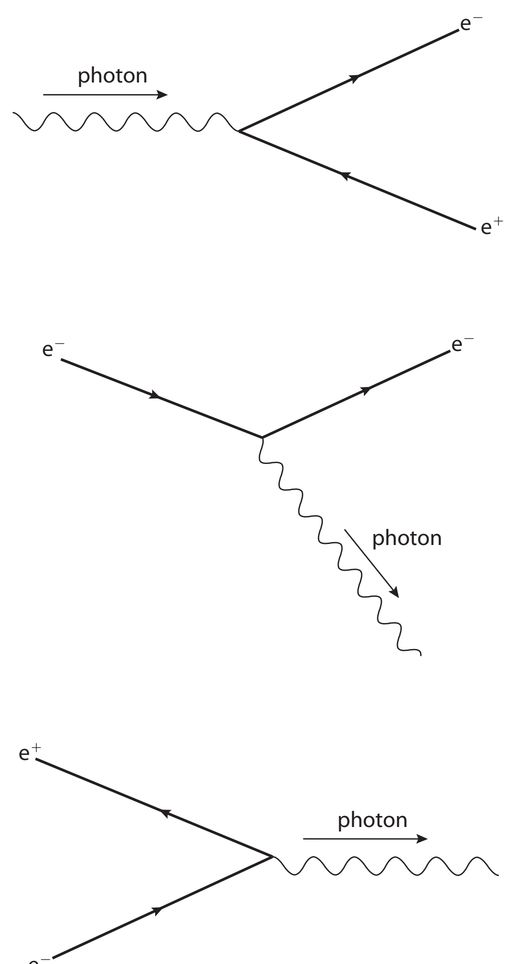


Figure 15.17 The first Feynman diagram shows an electron-positron pair production. Time progresses L to R so the arrows represent particles leaving. The arrow pointing to the vertex must therefore be an antiparticle. After an anticlockwise turn the diagram represents an electron emitting a photon. After a further turn the Feynman diagram represents an electron-positron annihilation. This time the arrow pointing away from the vertex is the anti-particle.

Is there really any point in predicting whether events that we will never be able to see can happen or not?

Exercises

33 Draw a Feynman diagram for:

a) a positron absorbing a photon

b) an electron absorbing a photon.

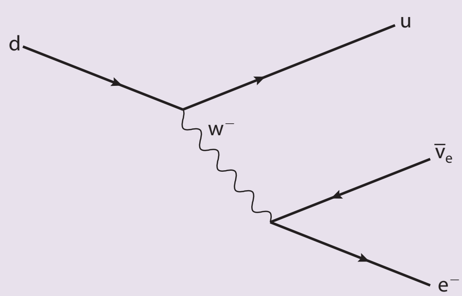
34 The following questions refer to the Feynman diagram below.

(a) Which particle comes into the interaction?

(b) What kind of interaction is this?

(c) What particles come out of the interaction?

(d) What well-known decay does this interaction represent?



35 By swinging the vertices around predict

(a) what would be emitted if a down quark interacted with a positron?

(b) if an anti up quark interacted with a positron what would be emitted?

15.7 Experimental evidence for the quark and standard models

Assessment statements

J.5.1 State what is meant by deep inelastic scattering.

J.5.2 Analyse the results of deep inelastic scattering experiments.

J.5.3 Describe what is meant by asymptotic freedom.

J.5.4 Describe what is meant by neutral current.

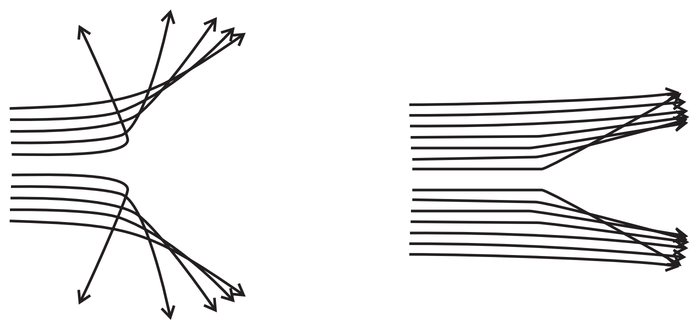
J.5.5 Describe how the existence of a neutral current is evidence for the standard model.

In the standard model, matter is broken down into the smallest number of fundamental units that are possible to construct all known particles. Using high-energy collisions it has been possible to support a lot of the predictions made by the theory, but due to quark confinement it is not possible to detect free quarks. However, by using high-energy electrons it is possible to show that the hadrons and mesons are made of smaller particles.

Scattering experiments: particle view

The principle of using scattered particles to investigate the internal structure of particles was used by Geiger and Marsden in their famous alpha scattering experiment. The idea is that it if you know the scattering angle of all the scattered particles, you can calculate what it was that scattered them. For example, four very small charges produce a very distinctive pattern as shown in Figure 15.18(a).

Figure 15.18 Scattering of alpha particles by four very close together positive charges.



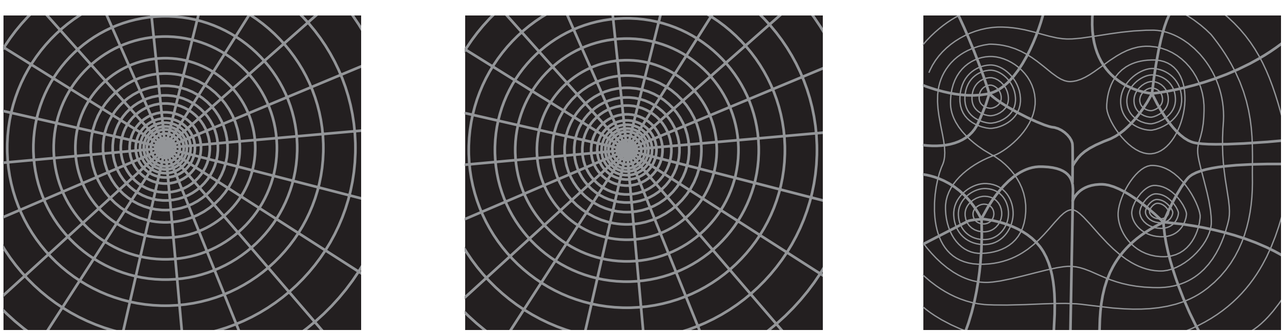
(a) Low energy

(b) High energy

However, we can't tell anything about the fine structure of the scattering object, because we don't get close enough. In fact you can't even see that there are four charges – it looks just like one. This is because they are so close together they look like one (we cannot resolve them).

To get closer we need to increase the energy of the scattered particles. Figure 15.18(b) shows the result of increasing the energy, but this is still not enough.

To understand what is happening we need to look at the electric field. When we are at a distance from a group of charges the resultant field is equal to the vector sum of the fields due to each individual charge. If we are much further away than the separation of the charges, then the resultant field is the same as if there were only one charge. This is shown in Figure 15.19.



(a) One particle, charge 4*q*

(b) Four particles, charge *q*

(c) close-up view of four particles

Figure 15.19 These pictures show the field pattern and lines of equipotential for one charge compared to four charges.

We can see that if we get very close to the charges, the field is very different and the scattering caused would now be much more complicated. However, if we get even closer, then the fields are again those due to simple point charges.

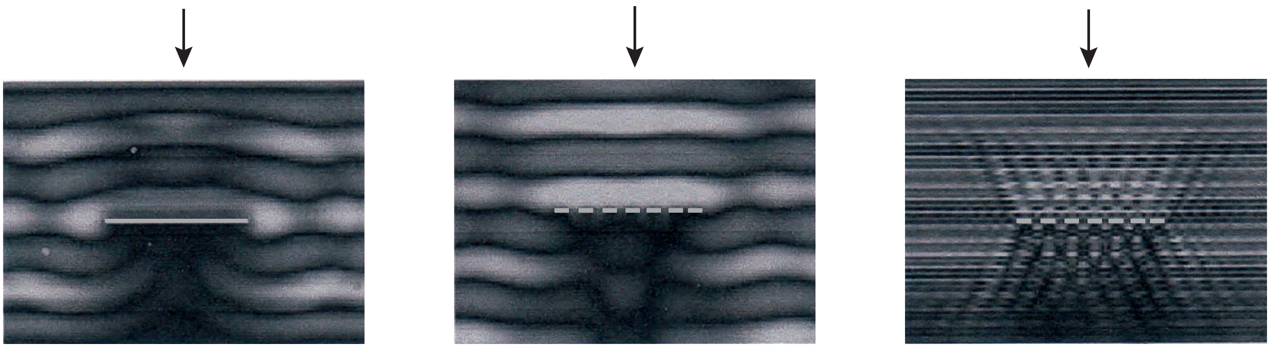
So if protons and neutrons are made of smaller particles then very high energy particles will be scattered by the field produced by these small particles. There is however one complication; when particles have such high energy they won't simply bounce off the targets elastically as in the alpha scattering experiments of Geiger and Marsden, they will knock the target particles out of position. This makes the analysis of scattering data more complicated but not impossible.

W

To access an applet to create field patterns such as the ones shown in Figure 15.19, visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 15.3.

Scattering experiments: wave view

When a wave passes a small object (as in figure 15.20A) diffraction occurs causing the wave to bend around the object. If the object contains holes that are much smaller than the wavelength (as in figure 15.20B) there is very little effect. However if the wavelength is reduced to approximately the same size as the holes then the pattern is quite different. So if we want to use the diffraction effect to know if a barrier has holes or not then we should use waves of short wavelength.



A B C

Figure 15.20 Waves of different wavelengths are diffracted by different objects.

Comparing A and B we can see that the small holes in B have made little difference to the diffraction pattern caused by the grey object. However, if we reduce the wavelength to the same size as the holes then a completely different pattern emerges.

We know that all particles with momentum have a de Broglie wavelength, λ, given by the equation $\lambda = \frac{h}{p}$. Electrons of energy in the 500 eV range have a wavelength about the size of an atom, so are diffracted by atoms. If electrons with GeV energies are used they have such small wavelength that they will be diffracted by the individual constituents of the nucleons. So by analysing the diffraction of high speed electrons we can find information about the internal structure of nucleons.

Deep inelastic scattering

Deep inelastic scattering is the scattering of high-energy particles by the particles that make up nucleons. This is inelastic since, during the process, the scattered particle loses energy to the target. By analysing the scattering angles and the product of such high-energy collisions it is possible to show that:

- Nuclei are made of smaller point charges that are knocked apart by the high-energy collisions. This is supported by the detection of protons and neutrons emerging from the collisions.
- At even higher energies: baryons scatter electrons in a way that is consistent with them being made of three point charges, and mesons in a way consistent with being made of two point charges. Furthermore, the size of the charges is $\frac{1}{3}$e or $\frac{2}{3}$e.
- The total momentum of the quarks is more than expected, implying that there are other particles in the nucleus – gluons.

These findings support the quark model, but due to the confinement principle, no free quarks were knocked off in the process. However, the particles that emerged were consistent with the standard model.

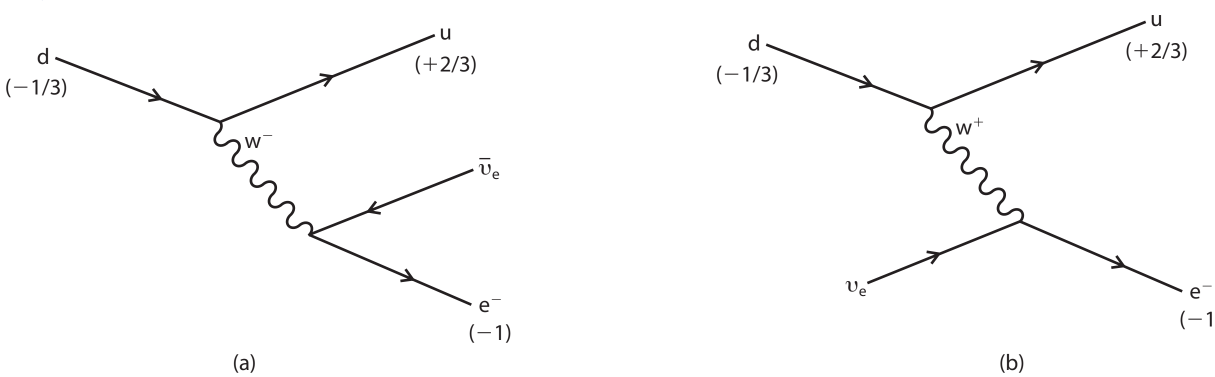
Asymptotic freedom

We know that the force holding quarks together is very strong and caused by the exchange of gluons. However, if the force increased as the distance between two quarks decreased, then matter would collapse to nothing. For that reason it is believed that at very short range the force must tend to zero; the quarks are said to have *asymptotic freedom* (you will have come across asymptotes in maths). If this is the case the quarks that make up a hadron will not be held tightly together by the strong force, but will be free to move about inside the particle like apples in a bag. It is getting the quarks out of the hadron that takes a lot of energy, not moving them around inside. By analysing the results of deep inelastic scattering it is possible to conclude that the small charges responsible for the scattering inside hadrons are in fact free to move around.

Neutral current

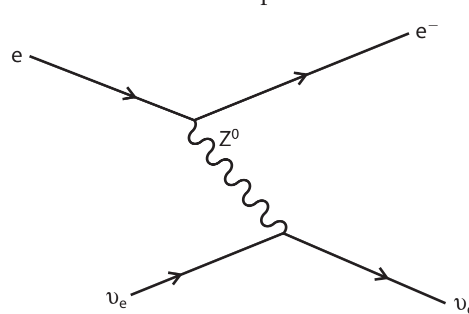
Applying the standard model we can predict that certain interactions are possible and others impossible. We have seen how Fey nman diagrams are used in this respect. For example if the interaction in Figure 15.21(a) is possible then interaction (b) must also be possible.

Figure 15.21 Feynman diagrams for two interactions. (a) is a beta decay and (b) is a prediction deduced by moving the vertices of the diagram around.



Since they involve neutrinos, both of these interactions are weak interactions; the exchange particle is therefore the W boson. These interactions also involve the moving around of charge (current). Following the rules of the standard model, the interaction in Figure 15.22 should also be possible.

Figure 15.22 The Feynman diagram for an interaction between a neutrino and an electron.



In this interaction there has been no transfer of charge, so the exchange boson is the Z^o. Interactions such as this are called *neutral currents*. Such interactions are very difficult to observe and were for a long time the missing link in the standard model. If they were not possible a 'no neutral current' clause would have to be written in to the model, which would have been untidy. In 1973 the first neutral current was detected in the Gagamelle bubble chamber at CERN.

The Higgs boson

The Higgs boson is included in this section about experimental evidence for the standard model because, if it is discovered, it will be evidence, and by the time you read this book, it could well have been discovered.

When looking at possible interactions, we always make sure that charge, lepton number etc. are conserved, but mass seems to come and go as it pleases. One explanation for this is that mass is not part of the particle but part of space called the *Higgs mechanism*. This is not so farfetched if we think of what mass actually is; it's the property of matter that makes something difficult to accelerate. Getting a heavy car to move along the road is more difficult than getting a football to move (even without friction). Why can't the property that makes it hard to move something be part of space like air resistance?

A common analogy used to describe how the Higgs mechanism might work is to think of a popular student entering a party. As she moves across the room, she is instantly surrounded by friends that make it difficult for her to proceed. She's perfectly OK if she stands still, but if she needs to go to the bar for a drink, she finds it very difficult to get going. At some time during the evening, another student comes into the room and tells a group of friends some gossip. This gossip quickly moves around the room as people cluster round to hear it. It is the cluster of students that make it hard to cross the room and this cluster can exist even without someone's progress to hinder.

For the Higgs mechanism to work, the whole of space must be covered by some sort of field (the Higgs field). When some particles move through this field they interact with it, causing them to exhibit the properties that we attribute to mass. In the same way that an electromagnetic field has a particle associated with it, the photon, the Higgs field is predicted to have a particle, the Higgs boson. (In the party analogy the particle is the cluster of students passing on the gossip.) It is this particle that is being hunted for.

15.8 Cosmology and strings

Assessment statements

J.6.1 State the order of magnitude of the temperature change of the universe since the Big Bang.

J.6.2 Solve problems involving particle interactions in the early universe.

J.6.3 State that the early universe contained almost equal numbers of particles and antiparticles.

J.6.4 Suggest a mechanism by which the predominance of matter over antimatter has occurred.

J.6.5 Describe qualitatively the theory of strings.

When particle physicists first realised that by looking at high-energy particles they could find out about the fine structure of matter, they had no particle accelerators, so they looked to the stars. Here, energies were high enough to split atoms into their constituent parts, so they could put together their theories by observing particles that arrived at the Earth. Today the situation is different; the collision energies available in the biggest reactors are even greater than on stars.

The Big Bang

To make a firecracker, gunpowder is compressed into a small cardboard cylinder. When ignited, a chemical reaction is initiated, which produces a lot of heat, increasing the KE of the gunpowder molecules. This increased KE results in increased pressure, enabling the gunpowder to push away the surrounding air in a rapid expansion. The idea of a Big Bang is that the universe was once very small, hot and expanding. There are three main reasons why this is thought to be the case.

Expanding universe

If the universe is expanding, then everything should be getting further apart. This would mean that all the stars we see in the sky would be moving away from us. If that were the case then the light we receive from them would have a longer wavelength (red shift). This is found to be true; however, it is important to realise that they are not flying apart through space – it is space itself that is expanding.

The Boltzmann equation
This equation relates the temperature of a gas with the average KE of the atoms.

$$KE = \tfrac{3}{2}kT$$

where
k = The Boltzmann constant $1.38 \times 10^{-23}\ \mathrm{J\,K^{-1}}$
T = temperature in kelvin.

7.7×10^{9} K is roughly equivalent to 1 MeV.

Cosmic background radiation

If there were a Big Bang, then soon afterwards the universe would have been an extremely hot, uniform and expanding mixture of radiation and matter. As the matter expanded, radiation would have been left travelling throughout the universe. This radiation should still be here, but due to the stretching of space, its wavelength will be much longer. In 1965 a uniform microwave radiation was detected without the strong spectral lines associated with stars. This suggests it is coming from something completely different, the Big Bang.

The abundance of helium and deuterium

In the chapter on nuclear physics you will have learnt about binding energy and fusion. We know that it is possible to join small nuclei together to form larger ones, but this requires energy to get the nuclei close enough together. This happens when nuclei fuse in stars, but there is far more helium and deuterium present in the universe than can be accounted for, if it were only being produced in stars. However, the amount is in agreement with the theory that the universe was once very hot and dense.

Before the Big Bang

At the beginning of this book the fundamental quantities of position and time that are needed to model the physical world were introduced. Without these quantities, we can't do physics. If we go back in time to less than 10^{-43} s after the Big Bang (the Planck time) there was no length and time. These dimensions are part of the universe, which hadn't been created yet. This makes it very difficult to use physics to explain what was there, let alone before the Big Bang itself. So it is probably best to simply answer 'we don't know'.

Why do we have to know everything?

Isn't it best that some things are left unknown?

Times and temps

We know from the study of matter that it is made of a number of fundamental particles arranged in different ways. According to the Big Bang theory, at the beginning of time, these fundamental particles were mixed up randomly, then, as the universe cooled, they grouped together to form the atoms and particles that we have today.

From what we know about particle physics, certain events require certain energies. For example, to knock electrons off an atom requires energies of a few eV but to knock protons out of a nucleus requires energies in the MeV region. From this information, we can work out the temperature of the universe when different events took place.

The beginning

We will start at 10^{-43} s after the Big Bang. At this time the temperature has been calculated to be 10^{32} K, that is equivalent to a particle energy of 10^{19} GeV. At these energies, matter as we know it could not exist; the universe was a very small sea of photons, leptons and quarks, equal numbers of particles and antiparticles. They would be moving too fast to combine, but as the universe expanded, it cooled.

10^{-35} s – The grand unification era

At 10^{-35} s the temperature had cooled to around 10^{27} K (10^{14} GeV) at this temperature, the particles and antiparticles combined to form radiation, and photons would be turning into particle-antiparticle pairs. These particles and antiparticles could then decay. At the end of this period more antiparticles had decayed then particles, resulting in a universe with more matter than antimatter.

10^{-12} s – The particle era

The temperature was now 10^{15} K (100 GeV). (We are now in the region of energies attained by a particle accelerator.) At this temperature the quarks start to form protons, neutrons and the other baryons and mesons that are around today.

3 minutes – The nuclear reaction era

At temperatures of 10^{9} K (0.1 MeV) protons and neutrons can combine to make nuclei. At the end of this era the universe is 90% hydrogen and 10% helium.

10^{5} years – The recombination era

Once the universe has cooled to below 4000 K (0.4 eV), electrons start to combine with nuclei to form atoms.

10^{9} years – The galaxy era

Gravity pulls large masses of atoms together, where they can combine to form the larger elements.

Now

The average temperature is now 3 K (0.0003 eV). There are still a lot of hot places (stars) where nuclear reactions are going on, but most of the universe is cold.

Exercises

36 At CERN it is possible to create collisions with an energy in the region of TeV. What temperature is that equivalent to?

37 Calculate the average KE in joules of the universe 10^{-12} s after the Big Bang. Convert this to eV.

38 The binding energy of helium is 28 MeV. Use this data to show why helium could not exist until 3 minutes after the Big Bang.

String theory

Whilst dealing with subatomic particles we have ignored gravity. This is because the mass of the particles is so small that the size of the force is negligible. Remember:

$$F = G\frac{Mm}{r^2}$$

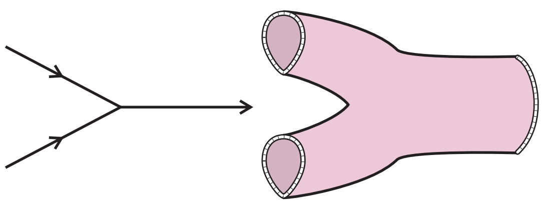


Figure 15.23 Two particles interact at a point but loops of string interact over a region.

But if the distance between the particles is also very small then the force is no longer negligible but very large. That's why we ignored gravity; it doesn't fit in with the idea of matter being made up of very small particles. String theory is an attempt to address this problem, by modelling fundamental particles as short strings (about 10^{-35} m long) instead of points. If we represent a particle by a loop of string then when two particles interact they do not have to get so close, as seen in Figure 15.23.

This is much more in tune with quantum mechanics, which tells us that the position of a particle is defined by a probability function. We can also model different particles by thinking of them as vibration in the string. So in the same way as there are only certain harmonics allowed in a given violin string, there are only certain particles allowed in the universe. To model particle properties with strings requires the use of extra dimensions, which is not a problem mathematically but makes them rather difficult to visualize.

W For more about quarks, string theory etc., visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink 15.4.

Practice questions

1 This question is about particle physics.

(a) Possible particle reactions are given below. They cannot take place because they violate one or more conservation laws. For each reaction identify one conservation law that is violated.

(i) $\mu^- \rightarrow e^- + \gamma$

Conservation law: (1)

(ii) $p + n \rightarrow p + \pi^0$

Conservation law: (1)

(iii) $\mathrm{p} \rightarrow \pi^+ + \pi^-$

Conservation law: (1)

(b) State the name of the exchange particle(s) involved in the strong interaction. (1)

(Total 4 marks)

2 This question is about deducing the quark structure of a nuclear particle.
When a K⁻ meson collides with a proton, the following reaction can take place.

$$\mathrm{K}^- + p \rightarrow \mathrm{K}^0 + \mathrm{K}^+ + \mathrm{X}$$

X is a particle whose quark structure is to be determined.
The quark structure of mesons is given below.

Particle	Quark structure
K^-	$s\bar{u}$
K^+	$u\bar{s}$
K^0	$d\bar{s}$

(a) State and explain whether the original K^- particle is a hadron, a lepton or an exchange particle. (2)

(b) State the quark structure of the proton. (2)

(c) The quark structure of particle 3 is *sss*. Show that the reaction is consistent with the theory that hadrons are composed of quarks. (2)

(Total 6 marks)

3 This question is about fundamental particles and conservation laws.
Nucleons are considered to be made of quarks.

(a) State the name of

(i) the force (interaction) between quarks. (1)

(ii) the particle that gives rise to the force between quarks. (1)

(b) Outline in terms of conservations laws, why the interaction $\overline{v} + p = n + e^+$ is observed but the interaction $v + p = n + e^+$ has never been observed. *(You may assume that mass-energy and momentum are conserved in both interactions.)* (3)

(Total 5 marks)

4 **(a)** Explain what is meant by a virtual particle. (2)

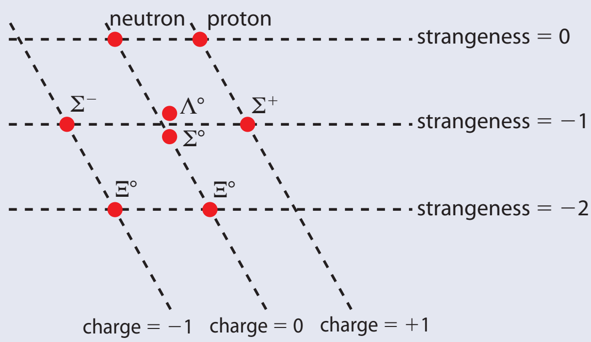
(b) Describe how the electrostatic repulsion between two like charges (Coulomb force) can be explained in terms of the exchange of virtual photons. (3)

(c) Complete the following table, listing the **four** fundamental forces, giving the exchange particles and characterising the ranges. (5)

	Force	**Exchange particle** (infinite or short range)	**Range of force**
1.	....	graviton	....
2.	Electromagnetic	photon	....
3.	....	$W^{\pm}, Z^0$	....
4.	....	....	....

(Total 10 marks)

5 One example of the success of organisational principles in science is Gell-Mann's 'eight fold way' which led to the development of quark theory.
The diagram below shows how eight baryons (each of **baryon number** = +1) can be grouped together according to **charge** and **strangeness**.



According to quark theory, each of these baryons are made up of three quarks. The properties of some quarks are given in the table below. Other types of quarks are known to exist.

Type of quark	Charge	Strangeness	Baryon number
u	$+\frac{2}{3}$	0	$+\frac{1}{3}$
d	$-\frac{1}{3}$	0	$+\frac{1}{3}$
s	$-\frac{1}{3}$	-1	$+\frac{1}{3}$
$\bar{u}$	$-\frac{2}{3}$	0	$-\frac{1}{3}$
$\bar{d}$	$+\frac{1}{3}$	0	$-\frac{1}{3}$
$\bar{s}$	$+\frac{1}{3}$	$+1$	$-\frac{1}{3}$

(a) What is the quark composition of the Σ^+ particle? Explain your answer. (3)

(b) Name the interaction and the associated exchange particle that binds these quarks together in the Σ^+ particle. (2)

(c) Name the three other types of quark that are known to exist. (1)

(Total 6 marks)

6 This question is about the conservation laws that govern the production, decay and interactions of fundamental particles.

Use the data in the table below to answer the following questions.

Particle		Mass ($MeVc^{-2}$)	Charge (C)	Baryon number	Lepton number
Neutron	(n)	939.6	0	+1	0
Proton	(p^+)	938.3	+1	+1	0
Antiproton	(p^-)	938.3	−1	−1	0
Electron	(e^-)	0.511	−1	0	+1
Antielectron	(e^+)	0.511	+1	0	−1
Pion	(π^+)	139.6	+1	0	0
Pion	(π^-)	139.6	−1	0	0
Lambda	(Λ^0)	1116	0	+1	0
Neutrino	(ν)	0	0	0	+1
Antineutrino	($\bar{\nu}$)	0	0	0	−1
Gamma photon	(γ)	0	0	0	0

The decay processes given below **do not** occur in nature. Determine **and** list the conservation laws that are violated in these processes. For each suggest a possible correct decay / interaction process.

Assume that the decaying / interacting particles are initially at rest.

(i) Neutron decay: $n \rightarrow p^+ + e^-$.

Does not occur because:

Process which does occur is: $n \rightarrow$ (2)

(ii) Lambda decay: $\Lambda^0 \rightarrow p^- + \pi^+$.

Does not occur because:

Process which does occur is: $\Lambda^0 \rightarrow$ (2)

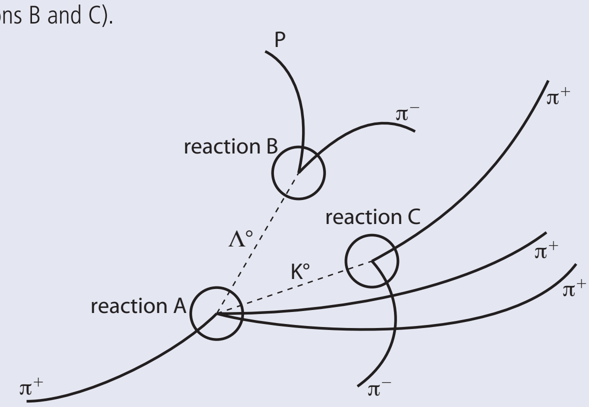
(iii) Electron annihilates with a positron: $e^- + e^+ \rightarrow \gamma$.

Does not occur because:

Process which does occur is: $e^- + e^+ \rightarrow$. (2)

(Total 6 marks)

7 The diagram shows the main features of a bubble chamber photograph in which a pion has collided with a stationary proton (reaction A). followed by two subsequent decays (reactions B and C).



The following information may be useful:

particle	baryon number	strangeness
π	0	0
$\Lambda°$	1	−1
p	1	0
K°	0	1

(a) Write an equation for **reaction A**. (1)

(b) Show that charge, baryon number and strangeness are all conserved in **reaction A**. (3)

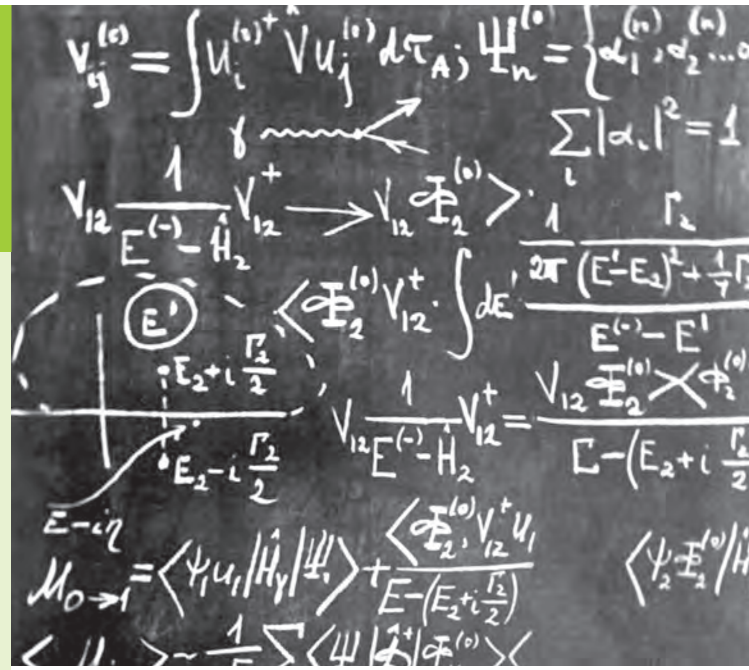
(c) How is it possible for there to be more pions at the end than at the beginning of the reactions? Where did they come from? (2)

(d) In addition to the quantities mentioned above, what else must be conserved in all three reactions? (1)

(Total 7 marks)

16 Theory of knowledge

"The task is not to see what has never been seen before, but to think what has never been thought before about what you see everyday."
Erwin Schrodinger



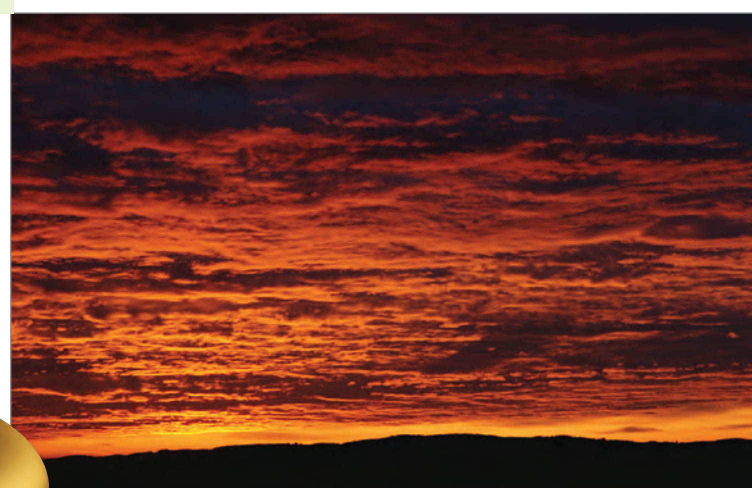
Are these particle physics equations beautiful?
Can you use your imagination to work out what they represent?

Introduction

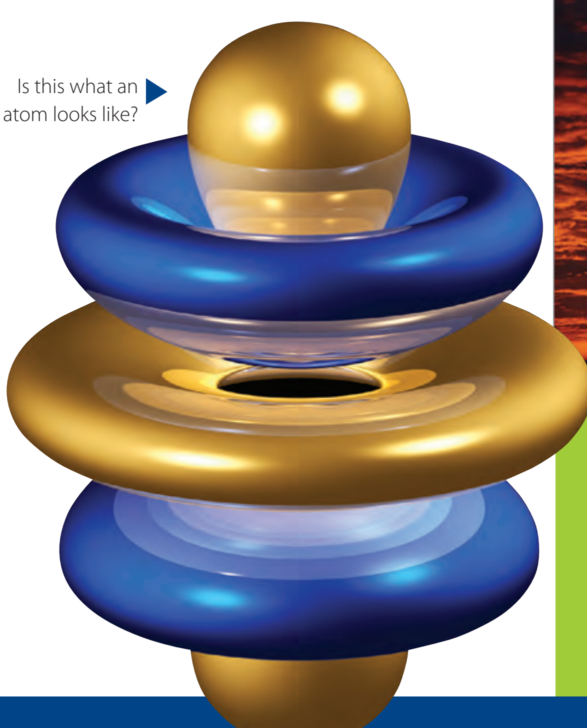
In the *Theory of knowledge* course you will be asked to analyse and discuss the different ways of knowing and areas of knowledge. Physics is one of the areas of knowledge that you have been introduced to, but what makes it different from other subjects such as art or languages? In this chapter we will look at the way that you gain knowledge in physics, so that you can compare this with the ways of knowing in your other subjects.

What is the role of imagination in physics? Can physics be beautiful? Is physics all logic and maths or is there a place for feelings?

Is this sunset more beautiful to people who understand Einstein's theory of general relativity?



Is this what an atom looks like?



"If we knew what it was we were doing, it would not be called research, would it?"
Albert Einstein

"The secret of genius is to carry the spirit of childhood into maturity."
T. H. Huxley

The scientific method

The scientific method is the way that physicists work to invent new theories and to discover new laws, and it is also the way that you will have been working in the practical program. There are actually many variations to this process and many exceptions, where new theories have come about without following any strict procedure. However, to make things simple, we will consider just one four-step version.

1 Observation

Physics is all about making models to help us understand the universe. Before we can make a model, we must observe what is happening. In physics, the observations are often of the form "How does one thing affect another?"

Franklin's lightning experiment. If Franklin had had a full understanding of electricity, he wouldn't have done this. He was lucky that the kite did not get struck by lightning, and in fact many people died repeating this experiment.

Galileo makes observations using a telescope.

2 Hypothesis

Having made an observation, the next step is to use your knowledge to develop an idea of what is causing the event you have observed. What factors cause this thing to happen and what factors are not involved? Having made a hypothesis, it is possible to predict the outcome of a change in one of the variables.

3 Experiment

All experiments are conducted to test hypotheses.

The experiment is designed to test the hypothesis. It is important to change only the quantity that you think is responsible for the event – you must keep the other variables constant.

If the experiment does not prove the hypothesis then you must go back to the observations and think of a new one (this is often helped by the outcome of the experiment). If the experiment supports the hypothesis then you can go on to the next step.

4 Theory

If experiment supports the hypothesis, you can make a theory that relates the variables involved. A theory is a set of related statements that can be used to make predictions and explain observations.

Example – the simple pendulum

Observation

A student watches a simple pendulum swinging and wonders what factors affect the frequency of the swing.

Hypothesis

The student hadn't studied the motion of the pendulum so used his intuition to come up with an idea related to the mass of the bob. He thought that since the bigger mass had more weight, then the force pulling it down would be greater, causing it to swing faster. So the hypothesis was that the frequency of the bob was proportional to the mass of the bob.

Experiment

An experiment was carried out measuring the frequency of bobs with different mass. The length of the string, the height of release and all other variables were kept constant. The result showed that there was no change; the hypothesis was therefore wrong.

Back to observations

On observing the pendulum further, the student noticed that if its length were increased, it appeared to swing more slowly. This led to a second hypothesis and the process continued.

"There are children playing in the street who could solve some of my top problems in physics, because they have modes of sensory perception that I lost long ago."
Robert Oppenheimer

Theories must be falsifiable

For a theory to be accepted it must be possible to think of a way that it can be proved wrong. For example: Newton's gravitational theory would be proved wrong if an object with mass was seen to be repelled from the Earth.

However, the theory that the Earth is inhabited by invisible creatures with eyes on each finger is not falsifiable since you cannot see the creatures to tell if they have eyes on their fingers or not.

Occam's razor

A razor is a strange name for a principle. Its name arises because it states that a theory should not contain any unnecessary assumptions – it should be *shaved* down to its bare essentials.

This is the same as the KIS principle "keep it simple".

For example, a theory for gravitational force could be that there is a force between all masses that is proportional to the product of their mass and is caused by invisible creatures with very long arms. The last bit about the invisible creatures is unnecessary so can be cut out of the theory (using Occam's razor).

"Just a theory"

If someone says "The theory of special relativity is just a theory", what do they mean?
The use of the word *theory* in the English language can cause some problems for scientists. The word is sometimes used to mean that something isn't based on fact. For instance, you could say that you have a theory as to why your friend was annoyed with you last night. In physics, a theory is based on strong experimental evidence.

Serendipity

Discoveries aren't always made by following a rigorous scientific method – sometimes luck plays a part. Serendipity is the act of finding something when you were looking for something else. For example, someone could be looking for their car keys but find their sunglasses. There are some famous examples of this in physics.

Hans Christian Oersted discovered the relationship between electricity and magnetism when he noticed the needle of a compass moving, during a lecture on electric current.

Arno A. Penzias and Robert Woodrow Wilson discovered radiation left over from the Big Bang, whilst measuring the microwave radiation from the Milky Way. They at first thought their big discovery was just annoying interference.

These serendipitous discoveries were all made by people who had enough knowledge to know that they had found something interesting. If they weren't expert physicists, would they have realized that they had discovered something new?

Previous knowledge

When you do practical work for the IB diploma, is the principle behind what you are doing the same as that used by leading physicists working in research departments of universities around the world?

What you are doing is using the knowledge learnt in class to develop your hypothesis. You should find out that, if you apply your knowledge correctly, your experiment will support your hypothesis. At the cutting edge of science, the scientists are developing new theories so the experiment is used to test the theory, not to test if they have applied accepted theory correctly. To find a new way of relating quantities requires imagination, but what you are being asked to do in your physics lab is to use accepted theory and not to use imagination. How can students trained to apply strict physical laws be expected to make imaginative new theories?

Hans Christian Oersted experimenting with magnets and current after a chance observation.

Paradigm shifts

A paradigm is a set of rules that make up a theory that is accepted by the scientific community. Having completed this course, you will have accepted certain paradigms. We see and interpret the world by virtue of paradigms and theories. Newtonian mechanics is a paradigm – we apply Newton's laws of motion to balls, electrons, planes and cars. The theory works well and is accepted by the scientific community. The way we treat almost everything as a particle is another paradigm. This paradigm is so entrenched in the way that we think that it is almost impossible to think of matter not being made of particles. How could you have a gas that was continuous? Before anyone thought of matter being made of particles, this wouldn't have been a problem, but now it is. To change your way of thinking requires a big leap of imagination, and this is called a paradigm shift. Throughout the development of physics there have been many paradigm shifts.

Copernicus

In 1543, when Copernicus suggested that the Sun was the centre of the solar system, it went against a theory that had been accepted for over a thousand years. Furthermore, it not only went against scientific theory it went against common sense, for how can we be going around the Sun when we are quite obviously standing still? It required a totally new way of thinking to accept this new idea. At the time the evidence was not strong enough to be convincing and the old paradigm remained. It was not until Galileo provided more evidence and later Newton developed an explanation, that the shift took place.

Einstein's theory of relativity

Before Einstein, it was accepted that time is the same everywhere, the length of a body is the same as measured by everyone and the mass of a body is constant. Einstein showed that time, length and mass all depend on the relative velocity of two observers. After this discovery it wasn't possible to carry on as if nothing had happened; what was required was a new way of thinking and a new set of laws. However, when a paradigm shift takes place, the old laws don't suddenly become obsolete – they just obtain limits. Newtonian mechanics is still fine when relative velocities are much less than the speed of light, and that is why it is still included in physics courses such as this.

The next paradigm shift

Will there be another paradigm shift? Can physics advance without one? Because there have been paradigm shifts in the past, does that mean that there has to be another one to advance physics further? In 1900 Lord Kelvin famously said "There is nothing new to be discovered in physics now, all that remains is more and more precise measurement." He was certainly wrong. Does that mean that if someone said the same thing today they'd also be wrong? One problem is that, as students go through the process of education, they get entrenched in the ways of thinking of their teachers, so the leap in imagination to make that new paradigm shift becomes bigger and bigger as time goes on.

Use of language

In physics, we use language in a very precise way; every time a quantity is named, it is given a specific definition. For example *velocity* means one thing and one thing only, the rate of change of displacement. In normal use of language, words can mean more than one thing; this is often used in jokes, poems and literature but not in physics.

"You feel awkward?
You're the one who said we ought to hunt bare."

Sometimes the other meanings of a word can lead to confusion. *Potential energy* sounds like a body could possibly have energy but it actually means the energy a body has owing to its position. *Electron spin* sounds like the electron is actually a little ball spinning; a spinning charge would indeed have the properties exhibited by the electron; however these properties do not arise because it's spinning.
To get around this problem, physicists sometimes use words that can't be confused with other meanings, who would think that a charm quark was actually charming, for instance?

The battleship and a bucket of water

Can a battleship float in a bucket of water? To solve this problem, you use Archimedes' principle that tells you that a body will float if it displaces it own weight of water. You might therefore conclude that the ship can't float in a bucket of water, because it has to displace (move out of the way) its own weight of water and there isn't that much water in the bucket. However, this is a misinterpretation of the theory. What Archimedes meant is that, when a boat is floating, if you filled the space in the water taken up by the boat with water, then it would be the same weight as the boat. This still might not make sense – that's why physicists use so many diagrams.

This demonstrates again how important it is to understand the language.

Opinions

If you see a painting, read a book or watch the news on the television, you will probably formulate some opinion about it. In many subjects that you study you are actively encouraged to formulate opinions and discuss them in class. For example, you could think the painting here is beautiful or you could think that it is horrible. Either way is fine, because it's your opinion and you can have whatever opinion you like when it comes to such things.

Can you have an opinion in physics? Is it OK to say that in your opinion Newton was wrong when he said that force was proportional to rate of change momentum and that you think that they are independent? In physics, opinions don't count for much, although they can sometimes be the beginning of the formulation of a testable hypothesis.

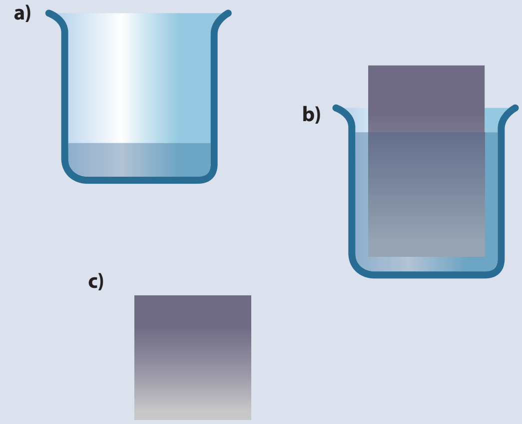


Figure 16.1 a) Bucket with a bit of water. b) Large object floating in bucket. The bucket is now almost full because the object is taking the place of the water. The object displaces the water. c) Archimedes says that the weight of fluid displaced equals the weight of the object. So this amount of water will have the same weight as the object.

"All truths are easy to understand once they are discovered; the point is to discover them."
Galileo Galilei

How much water do you need to float this huge oil tanker?

Laws

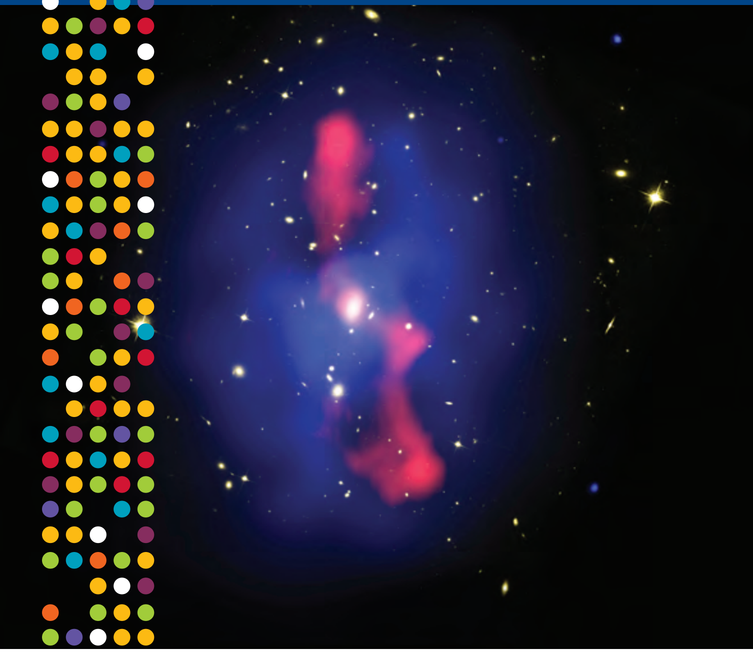
In Physics we use the term *law* quite a lot; for example, Newton's laws of motion, the law of conservation of energy and Ohm's law. The laws are generalized descriptions of observations that are used to solve problems and make predictions. If we want to know what height a ball will reach when thrown upwards, then we can use the law of conservation of energy to find the answer. When you use a law to solve a problem, you have a solid foundation for your solution. If someone were to disagree with your solution then they are disagreeing not only with you, but the law (assuming you applied it correctly). Laws sometimes give easy answers to difficult problems. If someone comes to you with a design of a machine that is 100% efficient, you don't need to study the details, because you can simply apply the second law of thermodynamics and say it won't work.

Universal laws

Some of the laws in physics are called universal laws; for example, Newton's universal law of gravitation. A universal law applies to the whole universe, but it is possibly naïve (or arrogant) to think that we can write laws that apply to the whole universe, when we can only make measurements from one very small part of it. Today, scientists are more modest in their claims and accept that there are probably parts of the universe that do not behave in the same way as things in our solar system.

Sense and perception

Physics is based on observation and observations are made with our senses. This was certainly true hundreds of years ago but today, although the information finally arrives into our brain via our senses, the observation itself is often done via some instrument. Copernicus had difficulty convincing anyone about his theory that the planets orbited the Sun, because he didn't have any convincing observations. He had predicted that Venus would have phases like the Moon, but couldn't observe this. Galileo used the telescope to observe that Venus did indeed have phases like the Moon. At first, this was still not accepted, since people didn't trust the telescope – they wanted to "see it with their own eyes".



Galaxy cluster MS0735.6+7421. The red part of this image is radio, the blue X-ray and the yellow is light. You could never see this.

Seeing is believing?

Can you see a hidden face in this picture? Apparently those with "physical brains" take a long time. See below for a hint.

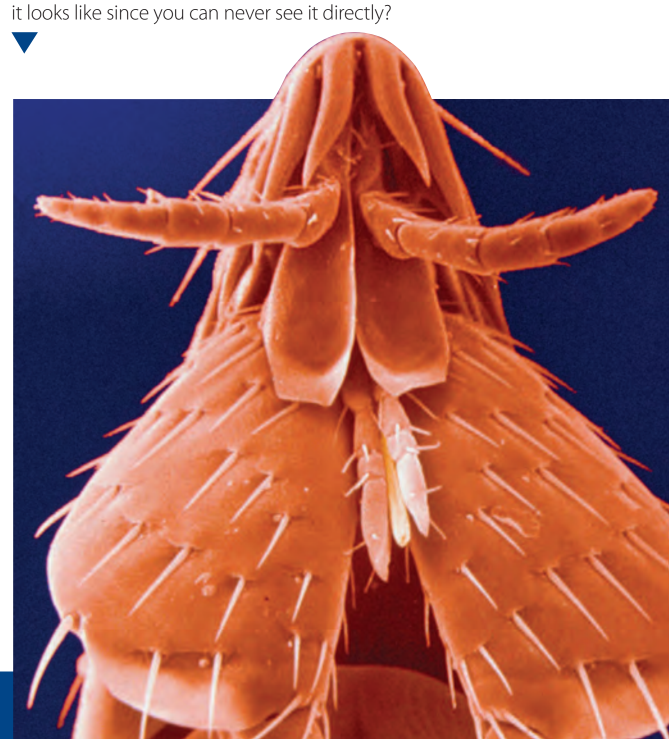
During the IB physics course you will have been asked to make observations and devise research questions; is this easier when you have studied the topic already or when it's something totally new? When you already know what you are supposed to be looking for, it is often easier to get started. However, if you have no preconceived ideas, you might have more chance of spotting something new.

Hint: look between 3 o'clock and 5 o'clock.

Deceptive pictures

A camera operates on the same physical principle as the human eye. Visible light is reflected from an object and focused by a lens onto a screen. It is reasonable to think that a picture is a good record of what we see. Using digital technology, it is now possible to recreate pictures from light that we can't see. Is this seeing? Can we say that we have seen a distant galaxy when we look at a picture constructed from radio waves? Can we say that we have seen the face of a flea when the picture was constructed from the diffraction pattern of electrons?

The face of a flea. Can you really say that this is what it looks like since you can never see it directly?



The Moon illusion

Have you ever noticed that the Moon looks bigger when it's close to the horizon? If you know some physics, you might have thought of an explanation. Maybe it's due to the refraction of the light by the atmosphere, or the elliptical orbit of the Moon causing it to sometimes be closer to the Earth. These theories would be based on the observation that you have made using your senses. However, if you were actually to measure the size of the Moon you'd find that it doesn't change as it moves from overhead to the horizon; it's an illusion. Physical theories are based on measurement not perception.

Imagination

Have you made use of your imagination during this course? Maybe not, but it is important that physicists have imagination, as Einstein famously said;

"I am enough of an artist to draw freely upon my imagination. Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world".

Without imagination, the huge leaps that have brought about the paradigm shifts in the way we perceive the world would never have been made. In learning basic physics, following rules tends to be emphasized more than using imagination, but at the cutting edge it is imagination that enables scientists to forge their way forward.

Children use their imagination all the time when playing. What happens to this skill in later life?

Physics or intuition?

One of the problems with studying physics is that we all live in the physical world and have all seen how bodies interact. We all know that if you drop something it falls to Earth and if you push someone on a swing they will move back and forth. These observations give us a feeling for what is going to happen in other instances; we call this feeling *intuition*, the ability to sense or know what is going to happen without reasoning. In physics, we create models to help give a reason for what is happening. This all works fine until intuition gives us a different answer to the laws of physics. Here's an example: Consider a metal bar floating in space, where the gravitational field strength is zero. If you apply two forces to the bar as shown in Figure 16.2, what happens?

Intuition will probably tell you that the bar will rotate about point A. This is because if you do this yourself that is what will happen. However that answer is wrong. Let's now apply Newton's laws of motion to the problem.

Newton's first law states that a body will remain at rest or with uniform motion in a straight line unless acted upon by an unbalanced force. The forces in this example are balanced, so the centre of mass of the bar (B) will not move.

We can see however that the turning effect (torque) of these forces is *not* balanced, so the rod will turn.

If the rod turns but point B doesn't move, then the rod must turn about B. And that is what happens.

The reason for the difference in these two predictions is that this rod is not in a gravitational field. When we try to do this with a rod on Earth, there are other forces acting.

Intuition was wrong; physics was right. The laws of physics can tell you what happens even if you can't do it or see it yourself.

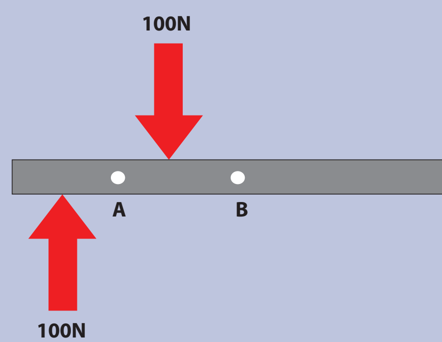


Figure 16.2 What happens when the bar experiences these forces?

Peer review

One of the strengths of modern scientific practice is that every new discovery goes through a rigorous process of peer review. Before a theory is published, it is sent to other scientists working in the same field. They give feedback to the research team before the theory is published. In this way mistakes can be spotted and problems ironed out. It also gives the possibility for other groups working in the same field to think of experiments that could be conducted to prove the theory wrong. Scientists are continually looking for ways to prove theories wrong, so when a theory is accepted by the scientific community one can be sure that it has been rigorously tested.

"It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong."
Richard Feynman

"There are many hypotheses in science which are wrong. That's perfectly all right; they're the aperture to finding out what's right."
Carl Sagan

"The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' (I found it!) but 'That's funny...' "
Isaac Asimov

Ethics and physics

Ethics is the study of right and wrong. It is sometimes not easy to decide when a course of action is right or wrong, and in these cases, it is useful to have a moral code or set of guidelines to refer to. In physics, there are two areas where ethical considerations are important;

1 The way physicists work in relation to other physicists; for example, they shouldn't copy each other's work or make up data.

2 The way their actions affect society; for example, physicists shouldn't work on projects that will endanger human life.

Whether a particular piece of research is ethical or not can be difficult to determine, especially when you do not know what the results of the experiment might lead to.

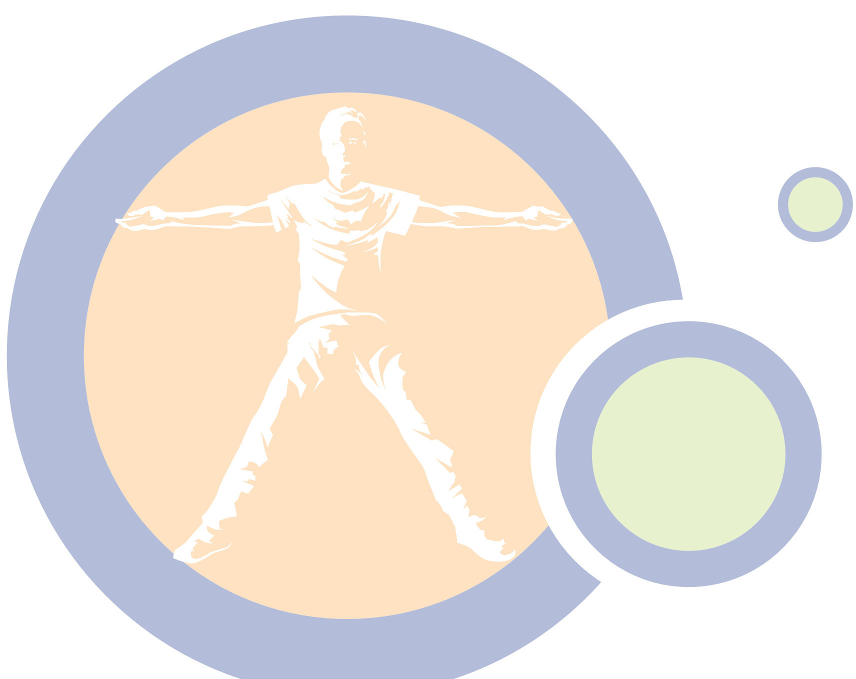
Should Rutherford have performed his experiments in nuclear physics, since the discovery of the nucleus led to the discovery of the atom bomb?

Can it be ethical to work in the weapons industry?

Who should decide whether a piece of research is carried out, physicists or governments?

If you left your body to science, would it be OK if it were used to test car seat belts? How about if it were used to test how far different types of bullets penetrate flesh?

Is it ethical to spend billions to carry out an experiment to test someone's hypothesis?



Internal assessment

During the HL IB physics course you should spend 60 hours doing practical work. During this time you will learn new skills and see how the theory can be applied in practice. Some of the reports that you write will be marked by your teacher according to specific IB criteria and will count for 24% of your final IB score. In this section, the criteria will be explained, so that you know what you have to do to achieve full marks.

There are five criteria: Design, Data Collection and Processing, Conclusion and Evaluation, Manipulative Skills and Personal Skills. Each criterion is split into three aspects, with each aspect graded on a scale from 0 to 2.

Complete – 2 marks

Partial – 1 mark

Not at all – 0 marks

Design

For most of the practical programme, your teacher will give you instructions on what you are going to measure and how you are going to measure it. As you become more familiar with the techniques, you will be asked to think of your own research question and devise a method for the collection of data.

For example, you could be given a rubber band with a mass tied to the end of it. Possible research questions could be:

- What is the relationship between the stretch of the rubber and the height the mass reaches?
- What is the relationship between the extension of the rubber and the height from which the mass is dropped?
- If I swing the mass in a circle, what is the relationship between the radius of the circle and the angular velocity?

Once you have decided on a research question, it must be stated clearly and the variables listed. These come into three categories:

- **Independent:** The one you are changing.
- **Dependent:** The one that changes because of the one you are changing.
- **Controlled:** Things that you don't want to change.

Once you have formulated your research question, you need to decide how you are going to measure the independent and dependent variables and how you are going to control those quantities that you do not want to change. It is always a good idea to include diagrams of your experimental set up and include details of how you make sure your measurements are accurate. At this stage you should also think carefully about the range of values and how many times you are going to repeat the measurements. Your report will then be assessed according to the following three aspects.

Aspect 1: Defining the problem and selecting the variables

IB rubric:

Complete	Formulates a focused problem/research question and identifies the relevant variables.
Partial	Formulates a problem/research question that is incomplete or identifies only some relevant variables.
Not at all	Does not identify a problem/research question and does not identify any relevant variables.

Checklist:

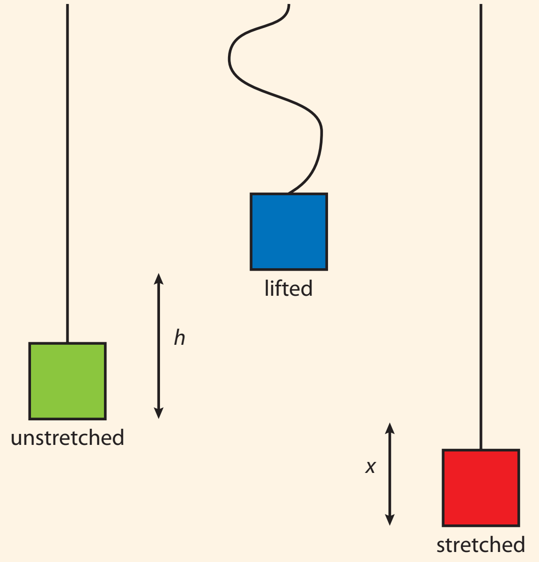
- State the research question clearly under the heading "Research question". It should be phrased in the form 'how is *y* dependent on *x*'. If the topic is not obvious, it is wise to write a paragraph introducing the topic before you state the research question.
- Identify and list the independent variable (this is the one you are changing, *x*) and dependent variable (the one that changes, *y*).
- Identify and list the controlled variables. These are all the other quantities that you could change but that are being kept constant.
- You will not be graded on writing a hypothesis but it is good practice to say what you expect to happen.

Report example

Introduction

This practical is an investigation into a rubber bung connected to an elastic band. The free end of the elastic band is clamped to a stand and the bung hung vertically from it. When the bung was lifted and released, the elastic band stretched (as shown in the diagram below). I decided to investigate the relationship between the maximum stretch of the elastic band and the height of release.

● **Hint:** Good idea to introduce topic since it's not obvious what this is about from the research question alone.



● **Hint:** Diagram helps clarify research question.

- **Hint:** Clear research question.
- **Hint:** Variables listed correctly.
- **Hint:** Controlled variables listed.
- **Hint:** Hypothesis included although not necessary for a complete score.

Research question

How does the extension of the elastic band (x) depend upon the height of release (h)?

Independent variable: The height of release

Dependent variable: The stretched length of the elastic

Controlled variables:

- The mass of the bung
- The length of the elastic band
- The type of elastic band
- The initial velocity of the bung

Hypothesis

Applying the law of conservation of energy, I expect that the gravitational PE at the top will equal the elastic PE at the bottom.

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

Since mg and k are constant I expect that x will be proportional to $\sqrt{h}$.

Aspect 2: Controlling variables

IB rubric:

Complete	Designs a method for the effective control of the variables.
Partial	Designs a method that makes some attempt to control the variables.
Not at all	Designs a method that does not control the variables.

To gain *Complete* in this aspect you should:

- List the apparatus used.
- Draw a labelled diagram of the apparatus (a photo is also a good idea).
- Describe how you are going to change and measure the independent variable.
- Describe how you are going to measure the dependent variable.
- Describe what you did to make sure the controlled variables remained constant.

- **Hint:** Apparatus list
- **Hint:** Details on how variables are varied and measured.

Method

Measuring the variables

<u>*Apparatus list*</u>

Plumb line

Ruler

Rubber bung

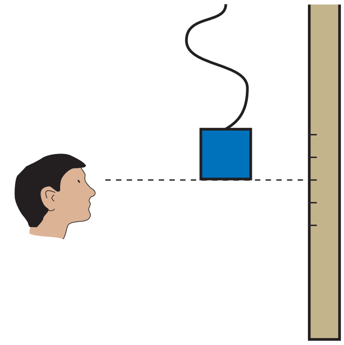
Elastic cord

To measure the height of release and extension, a ruler was mounted next to the elastic. It is important that the ruler is vertical so it was positioned using a plumb line.

All measurements were made from the bottom of the bung; I decided to do this because it was a straight line and therefore easy to line up with the ruler.

The bung was lifted so that it lined up with a cm mark on the ruler and released. To reduce parallax errors, I positioned my head in line with the bung when I took the reading. The ruler was positioned close to the bung but not touching.

After release, the lowest position of the bung was measured using the same ruler. I found that if I did this a couple of times I could position my head in line with the lowest point before release, again minimizing parallax error.



Controlling the controlled variables

The same bung and elastic band was used throughout the experiment.

After each run I waited a few seconds so that the elastic would lose any heat generated.

I was careful to make sure that the bung was released from rest each time.

● **Hint:** Details on how each of the controlled variables is kept constant.

Aspect 3: Developing a method for collection of data

IB rubric:

Complete	Develops a method that allows for the collection of sufficient relevant data.
Partial	Develops a method that allows for the collection of insufficient relevant data.
Not at all	Develops a method that does not allow for any relevant data to be collected.

Checklist:

- State the range of values of the independent variable that you are going to use.
- State how many times you are going to repeat the measurements of the dependent variable.

Report example

The experiment was repeated 5 times for each of 8 different heights ranging from 4cm above the "at rest" position to 12cm above. The elastic supplied by the teacher wasn't long enough to give the range that I wanted so I swapped it for a longer one.

I decided only to use initial positions where the elastic was slack. This is because I didn't want the elastic to have any elastic PE before release.

● **Hint:** A good range of values and each measurement repeated.

Data collection and processing

This criterion is all about how you construct your tables and process your data. You can do this by hand but it is easier if you process data using a spreadsheet programme such as Excel. To obtain a complete score in all aspects of this criterion, you will need to plot a straight line graph. This is not always possible when you have designed your own experiment but your teacher should give you plenty of examples where it is possible. When you are being assessed in this criterion, you will not be told what to do with the data. However, you will be expected to use a graphical method and that almost always means plotting a straight line graph and finding the gradient.

Aspect 1: Recording raw data

IB rubric:

Complete	Records appropriate quantitative and associated qualitative raw data, including units and uncertainties where relevant.
Partial	Records appropriate quantitative and associated qualitative raw data, but with some mistakes or omissions.
Not at all	Does not record any appropriate quantitative raw data or raw data is incomprehensible.

Checklist:

- Draw a table (using Excel) with a column for each measurement. This will generally mean one column for the independent variable and 5 for the repeated measurements of the dependent. There should be at least 5 rows, one for each time you change the independent variable.
- If your data is coming from the gradient of a 'data logger graph' or other graphic computer display, include an example of this graph in your report.
- The number of decimal places should be the same for all values in a column.
- Each column must have a heading and the units of the quantity.
- If you have estimated the uncertainty of the measuring instrument, this must be in the header. If you are going to calculate the uncertainty from the maximum and minimum values, you don't have to include it yet.
- Uncertainties should be rounded off to 1 significant figure, e.g. ±0.2 not ±0.17.
- The number of decimal places in the data should not exceed the limit of the uncertainty,
- e.g. if uncertainty is ±0.2 the measurement should only be quoted to 1 decimal place.
- Comment on how you arrived at any uncertainty value in the table.
- Comment on any observations you made that might be relevant later; there might not be anything here.

Report example

The data shown here is from an experiment to measure *g*.

Results

Raw data table

Below is a table of the data from the 5 runs performed for each of the 7 different heights.

● **Hint:** Errors and calculations explained.

The uncertainty in distance is estimated to be ±5mm due to the difficulty of measuring the position of the ball and the point at which the landing pad is activated.

Uncertainty in time is calculated from (Max Time – Min Time)/2

Distance (m) ±0.005m	Time 1 (s) ±0.001s	Time 2 (s) ±0.001s	Time 3 (s) ±0.001s	Time 4 (s) ±0.001s	Time 5 (s) ±0.001s	Av. Time (s)	Time unc. (s)
0.090	0.135	0.137	0.136	0.135	0.134	0.135	0.002
0.145	0.172	0.171	0.170	0.170	0.171	0.171	0.001
0.170	0.184	0.185	0.184	0.184	0.185	0.185	0.001
0.235	0.217	0.217	0.218	0.217	0.218	0.217	0.001
0.290	0.241	0.241	0.238	0.240	0.241	0.240	0.002
0.310	0.248	0.248	0.247	0.248	0.249	0.248	0.001
0.365	0.270	0.271	0.271	0.270	0.270	0.271	0.001

Measurements were taken from the bottom of the ball to the depressed landing pad.

Occasionally an obviously wrong value for time was obtained, but these were ignored.

● **Hint:** Table has consistent decimal places and units. Uncertainties seem reasonable.

● **Hint:** Good to make a note of any observations.

Aspect 2: Processing raw data

IB rubric:

Complete	Processes the quantitative raw data correctly.
Partial	Processes quantitative raw data, but with some mistakes and/or omissions.
Not at all	No processing of quantitative raw data is carried out or major mistakes are made in processing.

Checklist:

- The data should be processed in some way, for example averaging, squaring or finding the sine. Processed data should be displayed in a table separate from the raw data table.
- The table must have headers that include units and uncertainties.
- Calculate uncertainties in the repeated measurements by finding (max. value − min. value)/2 in the spread of data.
- Calculate the uncertainties in the processed data by calculating the (max. value − min. value)/2. e.g. if uncertainty in time is 0.2 then uncertainty in t^2 is $[(t + 0.2)^2 - (t - 0.2)^2]/2$.
- The number of decimal places in each column must be consistent with each other and the uncertainty.
- Any calculation must be explained.

$\frac{\text{Max.} - \text{Min.}}{2}$ will give an approximate value for the spread of the data. If you have a list of values you could find the standard deviation. However, for the IB, this approximate method is enough given the limited amount of data you are likely to record.

Hint: Table has consistent decimal places and uncertainties. All columns have correct units. Calculations explained.

Report example

Processed data

Since the initial velocity is zero, the vertical displacement and time are related by the equation $s = \frac{1}{2}at^2$ so a graph of s vs t^2 will give a straight line. The gradient of this line will be $\frac{1}{2}a$.

Distance (m) ±0.005m	Av. Time (s)	Time unc. (s)	Time^2 (s^2)	Unc. Time^2 (s^2)
0.090	0.135	0.002	0.0183	0.0004
0.145	0.171	0.001	0.0291	0.0003
0.170	0.185	0.001	0.0340	0.0002
0.235	0.217	0.001	0.0472	0.0004
0.290	0.240	0.002	0.0578	0.0007
0.310	0.248	0.001	0.0615	0.0005
0.365	0.271	0.001	0.0732	0.0003

The equation used to calculate the uncertainty in time^2 was (max. time^2 − min. time^2)/2 where the max. and min. values were taken to be the average value plus and minus the uncertainty.

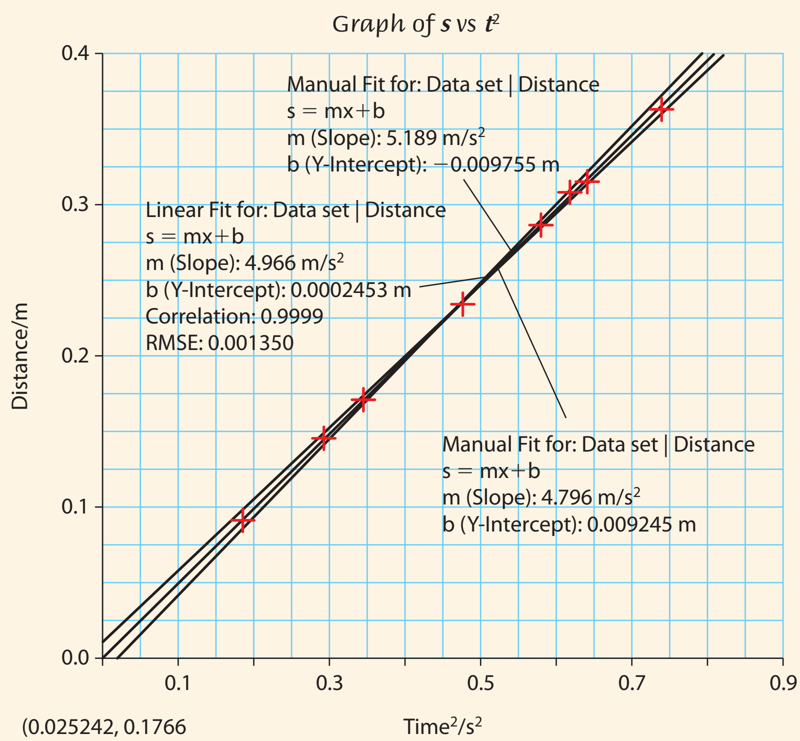
Aspect 3: Presenting processed data

IB rubric:

Complete	Presents processed data appropriately and, where relevant, includes errors and uncertainties.
Partial	Presents processed data appropriately, but with some mistakes and/or omissions.
Not at all	Presents processed data inappropriately or incomprehensibly.

- Processed data should be presented in a graph. This graph should be linearized if possible. The graph should be drawn using graphical analysis. *If not possible to linearize the function, a curve can be plotted. However, this makes the analysis more difficult, so the following points are for straight lines only.*
- The graph must have heading, axis labels and units.
- Independent variable should be on the x-axis
- Graph must include error bars.
- A best-fit line should be plotted automatically.
- The equation of the line must be displayed ($y = mx + c$).
- Manually fit the steepest and least steep lines that fit the error bars.
- The gradient should be displayed, including the uncertainty calculated from the steepest and least steep lines.

Report example



- **Hint:** Graph has correct labels, units, custom error bars, best-fit line, and maximum and minimum gradients.

$$\text{Max. gradient} = 5.189\,\text{m s}^{-2}$$

$$\text{Min. gradient} = 4.796\,\text{m s}^{-2}$$

$$\text{Uncertainty} = \frac{5.189 - 4.796}{2} = 0.197\,\text{m s}^{-2}$$

$$\text{Gradient} = 5.0 \pm 0.2\,\text{m s}^{-2}$$

Conclusion and evaluation

The conclusion of an experiment is what you found out from the results. This will usually include some value (e.g. the acceleration due to gravity) based on the gradient of a straight line. The most important thing to remember is that the conclusion must be based on your results, not on what you expected to happen.

In the evaluation, you comment on the precision and accuracy of your measurements, highlighting what was good and bad about your method.

Aspect 1: Concluding

IB rubric:

Complete	States a conclusion, with justification, based on a reasonable interpretation of the data.
Partial	States a conclusion based on a reasonable interpretation of the data.
Not at all	States no conclusion or the conclusion is based on an unreasonable interpretation of the data.

Checklist:

- State whether your graph supports the theory. For example, is the relationship between the quantities linear? This is only true if the line touches all error bars – don't say it is linear if it isn't.
- Are there any points on the graph that appear to be due to mistake (outliers)? Consider whether it's best to remove these and plot the line again.

In experiments you have designed yourself, you probably won't be looking for an actual value. Your conclusion will state whether or not your results agree with your hypothesis.

- Normally the data will be arranged so that the gradient will give you some value (e.g. 'g'). Calculate this value from the gradient.
- Calculate the uncertainty in this value from the steepest and least steep lines.
- Don't forget units.
- Compare your result with an accepted value, say where this value is from and quote uncertainty if known.

Report example

Conclusion

From the graph, it can be seen that within the uncertainties in the experiment s is proportional to t^2. Since the acceleration is therefore constant, we can apply the equation $s = \frac{1}{2}at^2$ so the gradient of the line can be deduced to be $\frac{1}{2}a$ where a is the acceleration of free fall.

From the graph, the gradient = $4.966\,\text{m}\,\text{s}^{-2}$ so the acceleration $g = 9.932\,\text{m}\,\text{s}^{-2}$

The uncertainty in the gradient can be found from the steepest and least steep lines.

Max. value = $2 \times 5.198 = 10.396\,\text{m}\,\text{s}^{-2}$

Min. value = $2 \times 7.796 = 9.593\,\text{m}\,\text{s}^{-2}$

Uncertainty = (max − min)/2 = $\pm 0.4\,\text{m}\,\text{s}^{-2}$

The final value obtained for g is therefore $9.9 \pm 0.4\,\text{m}\,\text{s}^{-2}$

The accepted value established by the Third General Conference on Weights and Measures is $9.80665\,\text{m}\,\text{s}^{-2}$. This lies within the limits of uncertainty of the experimental value obtained. However, it should be noted that this is a mean value, the value of g in the lab may not be exactly the same as this.

- **Hint:** Value of g calculated from the gradient. Uncertainty calculated from maximum and minimum lines. Values compared.

Aspect 2: Evaluating procedures

IB rubric:

Complete	Evaluates weaknesses and limitations.
Partial	Identifies some weaknesses and limitations, but the evaluation is weak or missing.
Not at all	Identifies irrelevant weaknesses and limitations.

Checklist:

- This is where you say if the conclusion is reasonable or not. You must have evidence for anything you write here. This can be from your results (the graph) or the observations you made during the experiment. You shouldn't say friction was a problem without evidence. It might help to do a small experiment to show that something was a problem.
- *Comments do not have to be negative.*
- Comment on whether your graph shows a trend; is it clearly a curve even

though the line passes through the error bars? Are the errors reasonable, are they obviously too big or too small?

- Comment on whether the intercept tells you anything. If it is supposed to be (0,0) and isn't, it might suggest a systematic error.
- Comment on whether you manage to keep the 'controlled variables' constant.
- Comment on the equipment used and the method in which you used it.
- Comment on the range of values and the number of repetitions.
- Comment on time management.

Report example

Evaluation

Looking at the graph I can see that the data points lie very close to the best-fit line, although there is some small deviation. The small error bars realistically reflect the accuracy of the measurement. The final value was quite close to the accepted value supporting this deduction.

Air resistance was not seen to be a problem; if there had been air resistance the graph would not have been a straight line.

Although the experiment gave a good value, the random uncertainty could be reduced by repeating the measurements more times or using a wider range of heights. In this case air resistance would start to be a problem so a smaller ball could be used.

The intercept was very close to the theoretical value of 0, this shows that the height measurement was carried out accurately with no zero error.

● **Hint:** Evaluation based on results. Error bars and intercept mentioned.

Aspect 3: Improving the investigation

IB rubric:

Complete	Suggests realistic improvements in respect of identified weaknesses and limitations.
Partial	Suggests only superficial improvements.
Not at all	Suggests unrealistic improvements

Checklist:

- List ways of improving the investigation (for example, reducing the uncertainties). Anything you write here must be related to something you mentioned in the evaluation. This in turn should be linked to the results. *Think like a detective, look for evidence.*
- If possible do a calculation or a small experiment to show how the improvement might improve the accuracy of the result.
- If you had more readings (wider range or more repetitions) would it improve your result?
- Is there any modification to the apparatus that would make the results better?
- If you made any modification to the original method then mention it here, as you will then get credit for suggesting improvements.

Report example

Improvements

The method gave good results but the uncertainty ±0.4 $m\,s^{-2}$ could be reduced. The weak point of the experiment was the positioning of the ball and the release mechanism. This was not completely stable which, coupled with uncertainty related to the exact position that the landing pad sensed the ball, meant we could only measure the height to ±5mm. The ball could easily move after the measurement; a more solid support would reduce this error.

To reduce the uncertainty in the height measurement, we would have to replace the ruler with something more accurate. Perhaps a vernier calliper could be used to position the ball. However, if the support was not made more stable, this would be pointless.

A bigger range of values is often seen as a way of reducing the uncertainty. However, if we dropped the ball from higher up, then air resistance may be a problem, since it is related to the speed of the ball, which would in this case be higher.

As stated earlier, there was no evidence that air resistance was a problem, probably due the short drops used. Repeating the experiment in a vacuum would therefore not lead to a significant improvement.

● **Hint:** All improvements supported by evidence either from the results or observation.

Manipulation

Manipulation refers to the ability to follow instructions and carry out techniques in a safe way. Your practical reports will not be assessed according to this criterion; it will be done by your teacher observing you as you work in the laboratory. At the end of the year, you will get a grade based on your performance throughout the course.

IB rubric:

	Aspect 1	Aspect 2	Aspect 3
Levels/ marks	Following instructions	Carrying out techniques	Working safely
Complete/2	Follows instructions accurately, adapting to new circumstances (seeking assistance when required).	Competent and methodical in the use of a range of techniques and equipment.	Pays attention to safety issues.
Partial/1	Follows instructions but requires assistance.	Usually competent and methodical in the use of a range of techniques and equipment.	Usually pays attention to safety issues.
Not at all/0	Rarely follows instructions or requires constant supervision.	Rarely competent and methodical in the use of a range of techniques and equipment.	Rarely pays attention to safety issues.

Personal skills

Personal skills are assessed only during the group 4 project. This is a 10-hour project where you work with students studying other science subjects. The idea is that you get a feel for how scientists from different disciplines collaborate in the real world. The project should include some international aspect, a consideration of ethics and the use of ICT. Different schools have different ways of organizing the project but all will assess personal skills according to the same criteria.

IB rubric:

	Aspect 1	Aspect 2	Aspect 3
Levels/ marks	Self-motivation and perseverance	Working within a team	Self-reflection
Complete/2	Approaches the project with self-motivation and follows it through to completion.	Collaborates and communicates in a group situation and integrates the views of others.	Shows a thorough awareness of their own strengths and weaknesses and gives thoughtful consideration to their learning experience.
Partial/1	Completes the project but sometimes lacks self-motivation.	Exchanges some views but requires guidance to collaborate with others.	Shows limited awareness of their own strengths and weaknesses and gives some consideration to their learning experience.
Not at all/0	Lacks perseverance and motivation.	Makes little or no attempt to collaborate in a group situation.	Shows no awareness of their own strengths and weaknesses and gives no consideration to their learning experience.

ICT and IB physics

During the practical course you are expected to use the following digital devices and computer software:

- Data logger
- Graph plotting software
- Spreadsheets
- Databases
- Simulations.

The minimum requirement is that you use each once. However, depending on the equipment available at your school, it could be much more. More details of how to use ICT in IB physics can be found on the website:
http://occ.ibo.org/ibis/occ/resources/ict_in_physics/

Extended essay

The extended essay is a 4000 word piece of independent research on an IB topic of your choice. Tackling an extended essay in physics can be a daunting prospect, but your physics teacher will supervise your research and will be on hand to give guidance and help solve any practical problems that you might come across. Your supervisor will also give you a booklet, 'The extended essay guide', giving guidance on how to construct the essay with some specific recommendations for physics.

Choosing a topic

If you have chosen a good topic, then writing an extended essay in physics can be quite straightforward. Your supervisor will help you but here are some additional guidelines:

- Don't be too ambitious; simple ideas often lead to the best essays. Students often don't believe that they can write 4000 words on something as simple as a ball of plasticine being dropped on the floor, but end up struggling to reduce the number of words.
- Make sure your topic is about physics. Avoid anything that overlaps with chemistry or biology and keep well away from metaphysics or bad science.
- Although the essay does not have to be something that has never been done before, it must not be something lifted straight from the syllabus.
- Avoid a purely theoretical essay unless you have specialist knowledge. The essay must include some personal input; this is very difficult if you write about some advanced topic like black holes or superstrings.
- It is best if you can do whatever experiments you require in the school laboratory under the supervision of your supervisor. If you do the experiments at home during the holiday, keep in contact with your supervisor, so your research is kept on the right track.
- Choose a topic that interests you; it will be easier to keep motivated when the going gets tough.
- Sports offer a wide range of interesting research questions but sometimes it is very difficult to perform experiments. Roberto Carlos' famous free kick is a fascinating topic for an extended essay, but not even he can do it every time, let alone with different amounts of spin. If you are keen to do this sort of research try to think how you can simplify the situation so it can be done in the laboratory not on the football pitch.
- You must not do anything dangerous or unethical.

The research question

Once a topic has been decided upon, you will have to think of a specific research question. This normally involves some experimental trials and book research. The title of the essay often poses a question that could be answered in many ways; the research question focuses on the way that you are going answer the question. It is important that as you write the essay you refer back to the original topic and don't get lost in the intricacies of your experimental method.

Examples of topics and research questions

Does the depth of a swimming pool affect the maximum speed achieved by a swimmer?

Rather than trying to measure the speed of swimmers in different depth pools, experiments were performed in the physics lab pulling a floating ball across a ripple tank. This led to the research question 'What is the relationship between the depth of water and the drag experienced by a body moving across the surface?'

Why isn't it possible to charge a balloon that isn't blown up?

This topic led to the research question 'What is the relationship between the electron affinity of rubber and the amount that the rubber is stretched'. To perform the experiment a machine was built that could rub different samples of stretched rubber in the same way.

Why does my motorbike lean to the left when I turn the handle bars to the right?

Rather than experimenting on a motorbike, experiments were performed in the lab with a simple gyroscope. The research question was 'How is the rate of precession of a spinning wheel related to the applied torque?'

Giovanni Braghieri (IB physics student and EE writer) riding his motorbike.

Performing the practical work

Most extended essays will involve some practical work and you should start this as early as possible. If it doesn't work or you find you don't have the right equipment, you might want to change the research question. You don't have to spend hours and hours on the experiment (although some students do). The whole essay is only supposed to take 40 hours, so keep things in perspective. Make sure the experiments are relevant to the research question and that you consider possible sources of error, as you would in any other piece of practical work. If you get stuck ask your supervisor for help. They can't do the essay for you but can help you solve problems.

Research

Remember that you're doing research, not a piece of internal assessment. This means that you should find out what other people have done and compare their findings with your own. This might be difficult if you have chosen a particularly novel topic but most things have been done before. You can try the internet but science journals found in university libraries are often the best source of good information.

Writing the essay

Once you have done some research and conducted your experiment, you are ready to write the essay. Remember you are trying to answer a research question, so get straight to the point. There is no need to tell a story such as how this has been your greatest interest since you were a small child. Make a plan of how you want your essay to be; the thread running through it is the research question - don't lose sight of this. Here is a plan of the essay mentioned above about the balloon:

- Introduction of the topic and research question - how the electron affinity of rubber is connected to the charging of a balloon.
- The theory of charging a balloon and electron affinity.
- Hypothesis based on the theory.
- How I am going to test the hypothesis.
- Details of experimental technique.
- Results of experiment.
- Interpretation of results including evaluation of method.
- Conclusion – how my results support my hypothesis and the findings of others.
- Why a balloon that is inflated cannot be charged.

What can go wrong

In the real world things are rarely as simple as they first appear and you might find that your data does not support you original hypothesis. This can be disappointing but shouldn't ruin your chances of writing a good essay. First make sure that you haven't made any mistakes in your initial assumptions or analysis of data, then try to think what why the experiment doesn't match the theory and write this in the conclusion. Don't pretend that it does match if it doesn't.

Extended essay assessment

The extended essay is marked by experienced physics teachers against 11 criteria. It is important that you understand the criteria, since if your essay does not satisfy them, it won't score well even if it is a good essay.

Research question (2 marks)

Most importantly, your research question must be physics and not just loosely related to physics. 'Did Isaac Newton's mother influence his laws of motion?' is not physics. 'An investigation into the relationship between the thickness of jelly and the attenuation of a laser beam' is. Assuming you have a good research question, make sure you emphasize it in the introduction of your essay; for example, the first paragraph.

Introduction

The introduction puts your research question in context; it is not supposed to be a story. Give some background information about the topic you are investigating to help the reader to understand the research question. For example, if your research question is 'The relationship between the velocity of a toy hedgehog and the angle of the slope', explain how the toy hedgehog works. However, don't include a story about the day you bought it and how your love of physics blossomed from that day forth.

Investigation

This mark is for the practical work that you carried out, or, in the case of a theoretical essay, the research. Include enough detail so that the reader can understand what you did, but do not get bogged down in detail. Remember it is an essay and not a lab report. Don't use titles such as 'Data collection and processing'. If you have used secondary data, make sure you reference the sources and give some indication of their reliability. If you have got the idea for your experimental design from a book or the internet, then quote the source. Make sure you estimate the uncertainties in all of your measurements and propagate them correctly through any calculations; all graphs should include error bars.

Knowledge and understanding of topic studied

To gain marks in this criterion, you must show that you understand the physics that you are using. This will be impossible if you don't understand it, so choose a topic that you either understand or think you will be able to understand if you read up on it. This is why it is not a good idea to write an essay on something like string theory. It is also why interesting applications of newtonian mechanics to novel situations often lead to good essays.

Reasoned argument

To have a reasoned argument that runs through an essay requires a good essay plan. When you have your data and know your conclusions, plan how you are going to tell the story. The introduction should lead into the experiment, the results should imply the conclusion, and the evaluation should be based on evidence that can be seen in the results. Essays in physics can become unconnected sections. Think carefully about how it fits together and if something takes you away from the main argument, leave it out.

Application of analytical and evaluative skills appropriate to the subject

Most essays in physics will include some mathematics. Make sure you understand what you are doing. Don't just copy derivations from a book or blindly use computer software. Analyse your data properly; some of the approximations for calculating errors used in the internal assessment are not good enough, if using large amounts of data. Evaluate your experimental technique honestly. Don't try to hide mistakes – it shows you understand what you are doing if you can spot mistakes.

Use of language appropriate to the subject

In physics, words don't have two meanings. Use the language of physics carefully. If you use symbols to represent quantities, define them clearly and be consistent. Always give the units of any quantity.

If you don't know what a term means then don't use it; stick to what you know.

Conclusion

The results of your experiment should lead logically to the conclusion; this is part of the development of the argument mentioned previously. When you first thought of your research question, you might already have thought of the conclusion. Try to forget this and base the conclusion on what your experiment tells you, not on what you thought would happen. Your conclusion will have greater validity if your uncertainties are small. If they are large, explain how they affect your conclusion. If your results are inconclusive say what further investigation could be done to resolve the problem.

Formal presentation

Make sure you have included all of the components listed in the official extended essay guide:

- Title page
- Abstract
- Table of contents
- Page numbers
- References
- Bibliography.

Abstract

The abstract is an overview of the whole essay, including the research question, method of research and conclusion. Here is an example:

The Relationship between the Depth and the Drag of Water

The aim of the essay is to investigate the relationship between the depth of water and the resisting force caused by the water on a floating object that is being pulled parallel to the surface of the water. The experiment only deals with a small spherical object that is being pulled with a constant force, on a low velocity and on shallow depths to limit the scope.

According to the developed hypothesis the resisting force, drag, is proportional to $1/\text{depth}^2$ because the movement of the sphere pushes the water towards the bottom which means that the bottom is also pushing the water towards the sphere. The longer the distance between the sphere and the bottom the more the force is dispersed to other directions.

A method of measuring the acceleration of the sphere at a certain velocity but different depths was used to examine the relationship. From the acceleration, the masses and the gravitational force involved it is possible to calculate the drag.

The conclusion of the experiment is that the hypothesis does hold true for the conducted experiment i.e. the drag is proportional to $1/\text{depth}^2$ for the limited scope of situation that the experiment deals with. There are also certain reservations about the accuracy of the experiment.

Joonas Govenius RCNUWC

Holistic judgement

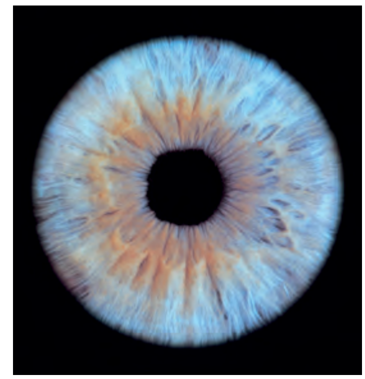
These marks are awarded to the essay as a whole to reward intellectual initiative, insight, originality and creativity. Even if the essay isn't well written, it can still gain marks here if you have, for example, shown original thought in devising an ingenious way of solving a practical problem.

Appendix (Option A)

The eye and sight

Assessment statements

A.1.1 Describe the basic structure of the human eye.
A.1.2 State and explain the process of depth of vision and accommodation.
A.1.3 State that the retina contains rods and cones, and describe the variation in density across the surface of the retina.
A.1.4 Describe the function of the rods and of the cones in photopic and scotopic vision.
A.1.5 Describe colour mixing of light by addition and subtraction.
A.1.6 Discuss the effect of light and dark, and colour, on the perception of objects.



The coloured part of the eye is called the *iris* and it consists of a ring of muscle fibres that contract and relax to alter the amount of light entering the eye through the pupil.

When the human eye is viewed from the front, the pupil looks black. In fact it is a circular hole in the centre of the iris, and the black colour comes from a layer inside the eye that prevents internal reflections. If the intensity of incident light is high, then the iris reduces the size of the pupil; in low intensity light, the pupil dilates to allow more light to enter.

The front of the eye is covered with a layer of transparent skin called the *cornea*. The lens and the cornea both play a part in focusing light onto the retina at the back of the eye. The retina is the light-sensitive layer and it sends messages to the brain through the optic nerve.

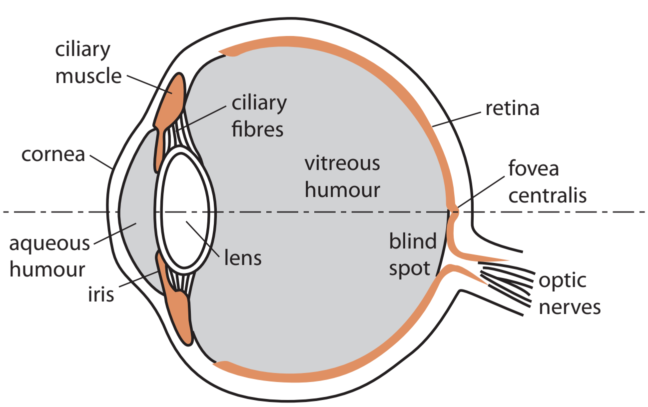


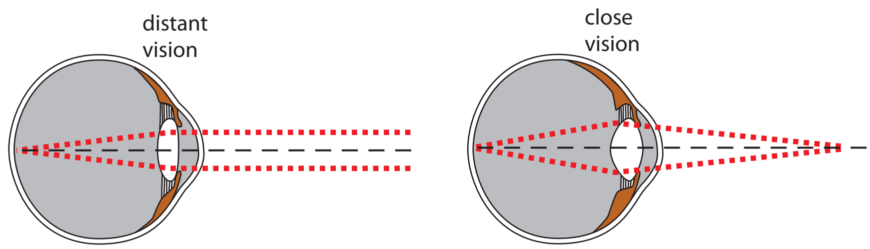
Figure 1 The human eye. The humours are transparent jellies that nourish the eye and keep its shape. The blind spot is where the optic nerve exits the eye and there is no retina. The fovea is directly opposite the centre of the lens and is sometimes called the yellow spot.

The lens in our eyes can change shape to enable us to focus on both near and far objects.

This process is called *accommodation*. When we look at a distant object, the ciliary muscles relax and the taut fibres make the lens longer and thinner. When we focus on something close to our eyes, the ciliary muscles contract and the lens changes shape to become shorter and thicker.

Modern laser surgery can reshape the cornea to correct many defects of vision and remove the need for glasses or contact lenses.

Figure 2 Rays of light from a distant object arrive at the eye effectively parallel, and so need less refraction than rays from an object close to the eye.



The near point is defined as the position where an object can be seen most clearly. As you are reading this page you are automatically holding it at your near point. If you hold it closer to your eyes it will appear blurred and if you hold it further away the writing will appear smaller and so more difficult to read.

The range over which the eye can focus is from the *near point* to the *far point*. For a normal eye, the far point is at infinity; we focus on the far point when the ciliary muscles are completely relaxed and the eye is not accommodating. The near point varies from person to person but is taken to be 25 centimetres.

Depth of vision

Depth of vision is the ability to see things in three dimensions. It is the perception of depth th at enables us to judge how far away things are, and this is crucially important when playing ball games or driving a car.

Using only one eye we can get some information about depth of vision. This is based on our previous experience and from cues we get from relative sizes of objects, perspective and relative motion between objects.

If the image from an object falls on the blind spot, it cannot be detected by the brain. With two eyes one compensates for the other, so we do not lose sight of things.

With two eyes we are able to obtain two images of the same thing and the brain is then able to judge distances and motion far more accurately. This is particularly important for nearby objects; for things that are far away there will not be a big difference in the images received by each eye.

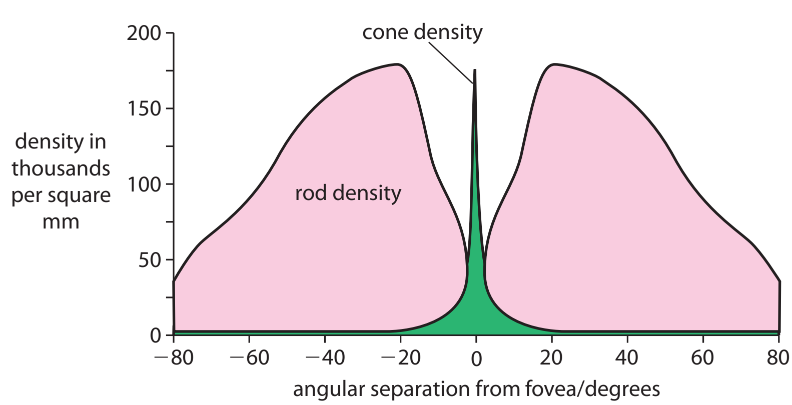
Colour vision

The retina contains two types of receptor cells called *rods* and *cones*. The rods detect motion, enable us to see in low light intensity and are responsible for peripheral vision (seeing things from the corner of our eyes). The cones are responsible for colour vision and also visual acuity. We need good visual acuity for reading the very small letters at the bottom of opticians' charts.

People in different parts of the world literally see things differently. The indigenous peoples of the Australian and Kalahari deserts have been found to have exceptional long distance vision; useful when scanning the horizon for moving animals. People who have spent their whole life in the rainforest have been found to lack depth perspective. When they see animals at a distance on the plains, they simply assume that the animals are very small; they have no other means of making a judgement.

The fovea or yellow spot, located in the centre of the retina, consists entirely of cones. This is the part of the retina where our vision is most acute. Just a few degrees from the fovea, the concentration of rods is at a maximum and the rods spread out all around the rest of the retina with a gradually decreasing concentration.

Figure 3 Variation in density of rods and cones across the retina. The diagram shows high density of cones in the fovea but low density elsewhere. There are no rods in the fovea but they are highly dense in other parts of the retina.



In each eye we have about 120 million rods but only about 6 million cones.

Experiments show that we have three types of cones in our retina. They are known as *blue*, *green* and *red* or as S, M and L which stands for short, medium and long, referring to the wavelength of light to which they are most sensitive.

There is overlap between the wavelengths to which each of the three types of cone responds and this enables the brain to perceive the full spectrum of colours. You can see from the diagram that both green and red cones respond to yellow light, but the blue cones do not respond to yellow or red light.

Photopic vision refers to colour vision under normal lighting and is the function of the cones. Scotopic vision refers to our ability to see in dim light, it is completely lacking in colour and is the function of the rods.

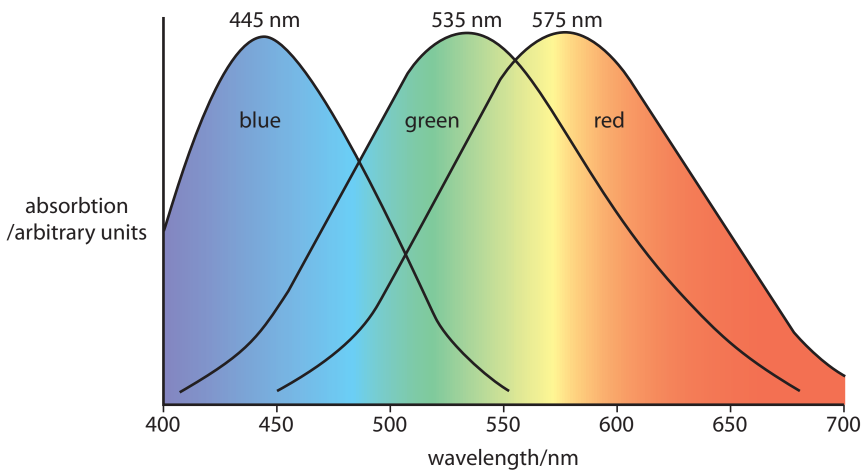


Figure 4 Spectral response curves for cones: each of the three types of cone responds to a different region of the visible spectrum.

Only 2% of our 6 million cones are for seeing blue, but since we appear to see equally well in the short end of the visible spectrum, the blue cones must be more sensitive.

Examiner's hint: You need to be able to sketch and interpret this type of graph.

People who are colour-blind have difficulty in distinguishing between different colours. The most common type is red-green colour-blindness and is inherited. It stems from problems in the green and red cones and involves an inability to properly distinguish between reds, yellows and greens. There are other types of colour blindness but they are quite rare.

Some apes also have been found to have three types of cones, so they see colours in a similar way to us. Other animals see things very differently; the vision of bees extends into the ultraviolet region of the electromagnetic spectrum, while some snakes can see infrared. Sharks are apparently unable to detect any colours.

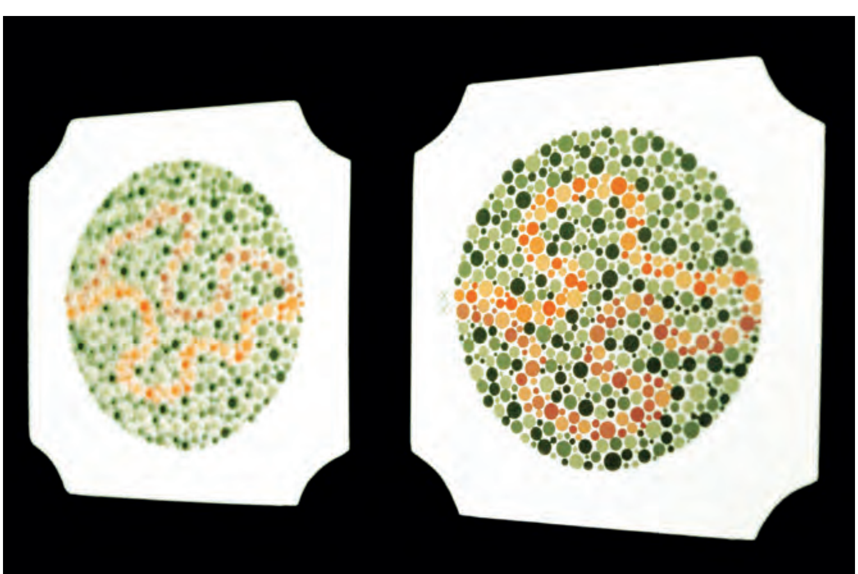
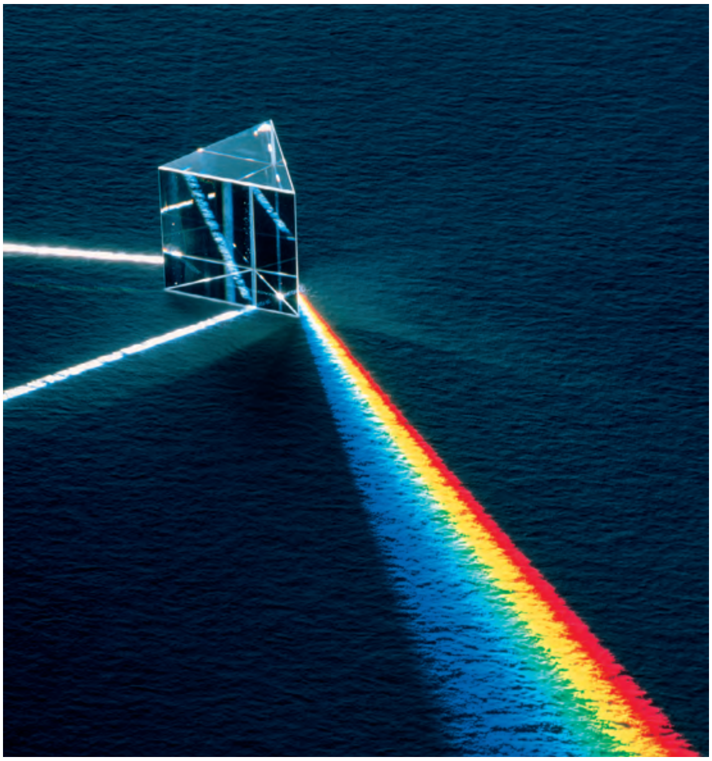


Figure 5 This is an example of a test used to see if people are colour-blind. A person with normal colour vision will be able to trace both the orange and red lines. A person who is colour-blind will find one line easier to follow than the other.



The white light is dispersed into the colours of the spectrum because different colours of light travel at different speeds in glass.

Colour mixing of light

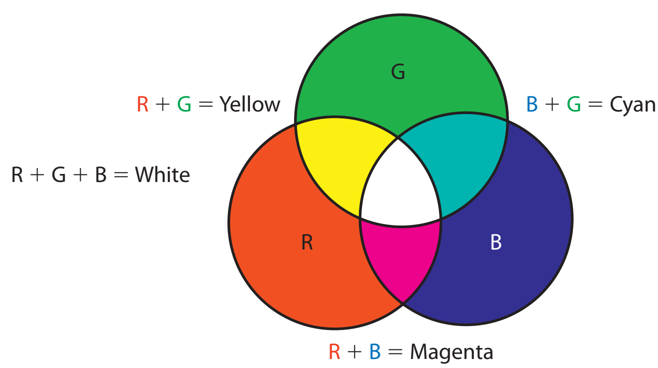
White light is a mixture of all the colours of the visible spectrum. A prism can separate the colours by a process called *dispersion*.

If three projectors are arranged to shine red, green and blue light of the correct intensity onto a screen, the result will appear white. It is for this reason that these colours are called *primary colours*.

Mixing the three primary colours in pairs gives us the three secondary colours.

Colours of light can be subtracted by absorption using filters. The colour of light that is transmitted by the filter is the same as the colour of the filter itself. For example, if white light is shone through a red filter, then all the wavelengths will be absorbed except for red; only red light will be transmitted.

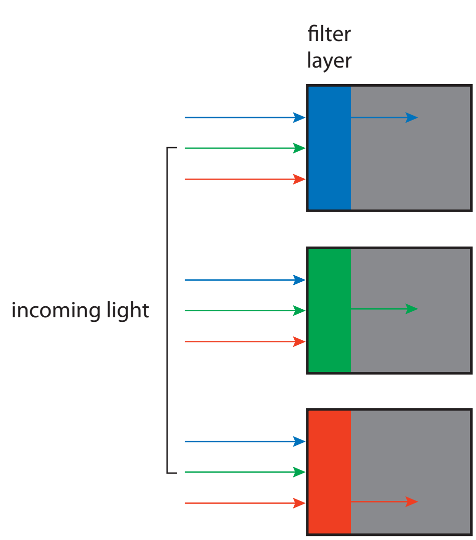
Figure 6 Adding red to green makes yellow. Adding red to blue makes magenta. Adding blue to green makes cyan. Adding all three together makes white. This is called additive colour mixing.



Red, green and blue are the primary colours of light.

Yellow, magenta and cyan are the secondary colours of light.

Figure 7 If a second green or blue filter is placed after the red filter, it will block the red light and then no light will be visible.



When white light passes through filters of the secondary colours, the wavelengths absorbed depend on the make up of the secondary colour. For example, magenta is made up of red and blue, so magenta transmits red and blue.

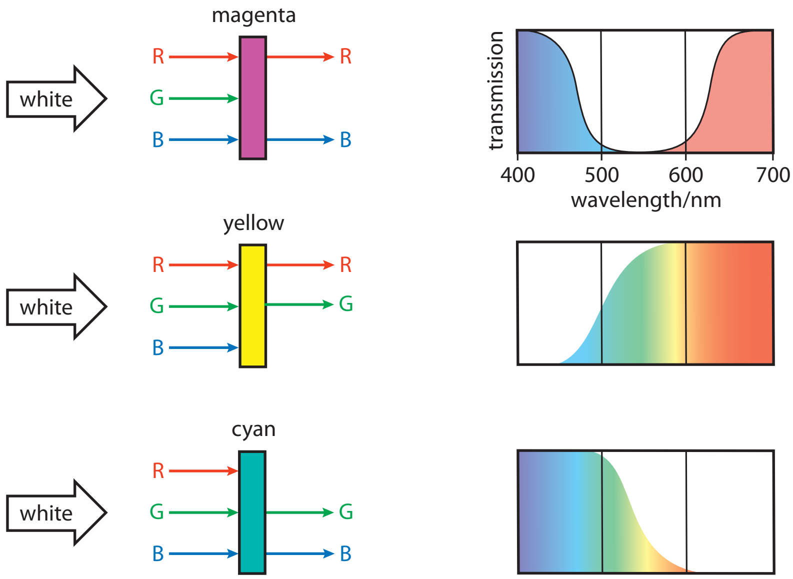


Figure 8 If another secondary filter is placed after the first one, then one primary colour will still be transmitted.

To view a fun simulation 'Color vision and filters', visit www.heinemann.co.uk/hotlinks, enter the express code 4426P and click on Weblink A.1.

Perception

Information collected by the eye is sent to the brain for processing. There is so much information coming in all the time that the brain must select what is relevant. Perception involves collection, selection and also organization and interpretation of the sensory input.

Light, shade and colour are used by artists, designers and architects to deliberately influence, and even alter, our perception of reality. For example, deep shadow in a painting, or in a church, can give the impression of massiveness.

Blue is perceived as a cold colour, orange and red glow can give the impression of warmth. A room can be made to appear bigger and lighter by the careful placing of mirrors and it can be made to appear higher by a light-coloured ceiling.

Sense perception is one of the ways of knowing. To what extent do our eyes give us knowledge of the world as it really is?

Exercises

1 Explain the process of accommodation in the human eye.

2 Distinguish between photopic and scotopic vision.

3 Sketch and label a graph to show the spectral response of cones in the human eye.

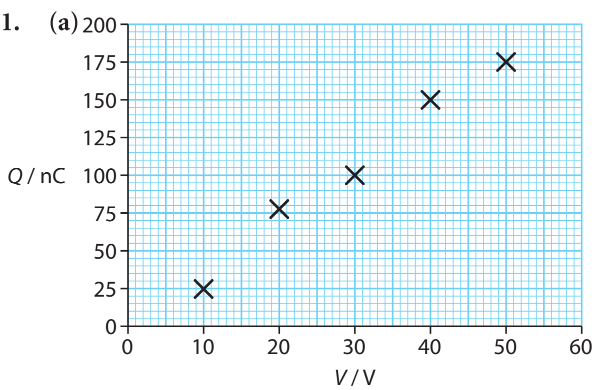
Answers

Chapter 1

1. (a) 5.585×10^6 m (b) 1.75 m
   (c) 2.54×10^{-5} m (d) 10^{26} m
2. (a) 2.69×10^9 s (b) 2.5×10^{-3} s
   (c) 3.46×10^5 s (d) 1.0379×10^4 s
3. (a) 0.2 kg (b) 1×10^{-8} kg
   (c) 2×10^3 kg
4. 150 m^3
5. (a) 1.0×10^{-10} m^3 (b) 1.09×10^{21} m^3
6. 180 kg
7. 86.85 kg
8. 5.48×10^3 kg m^{-3}
9. (a) 0.5% (b) 0.3%
   (c) 158.36 cm^2 (d) 0.8%
10. (a) 5.2 cm (b) 4.8 cm
    (c) 3 cm (d) 8.8 cm
11. (a) 5 cm (b) 5.66 cm
    (c) 6.32 cm (d) 3.6 cm
12. 8.94 km, 26.6°
13. 112 km, 26.6°
14. 8.66 km
15. 7.52 km
16. 433 m

Practice questions

1. (a)



half area of graph paper at least to be used;
axes labels including units;
scale;
data points; *((0, 0) need not be included)* 4

(b) absolute uncertainty in Q at 10.0 V = ±3 nC;
absolute uncertainty in Q at 50.0 V = ±18 nC;
Or read from graph or elsewhere in the question and do not deduct unit mark.
correct placing on graph; 3

(c) from top of error bar at (50, 180) to bottom of error bar at (10, 30);
use of at least half the line or algebraic indication;
value = 4.3 ***or*** 4.3×10^{-9}; 3
Watch for ecf.

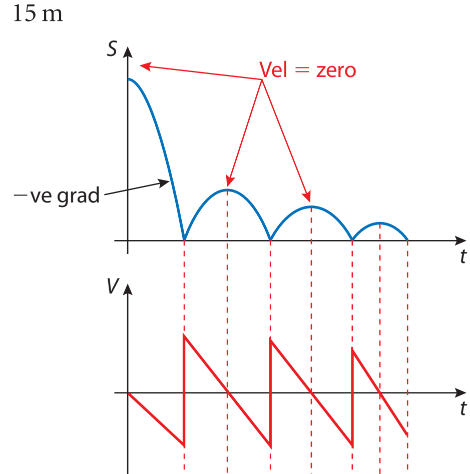
(d) CV^{-1}; 1
Unit might be given in (c).

(e) recognize that the gradient $m = \frac{\varepsilon_0 A}{d}$;
therefore $\varepsilon_0 = \frac{dm}{A}$;
$= \frac{0.51 \times 10^{-3} \times 4.3 \times 10^{-9}}{0.15}$;
$= 1.5 \times 10^{-11}$ C V^{-1} m^{-1} (C^2 N^{-1} m^{-2} – data book unit or F m^{-1}) ; 4

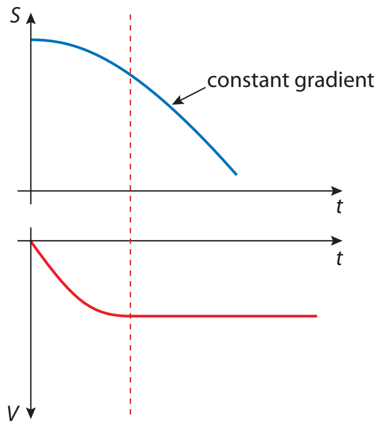
2. C
3. A
4. C
5. C
6. C
7. D
8. B
9. C

Chapter 2

1. (a) 27.8 m s^{-1} (b) 5.6 m s^{-1}
2. 22.4 m s^{-1}, 26.6°
3. 31.6 m s^{-1}
4. 50 m s^{-1}
5. −30 m s^{-1}
6. 40 m
7. 3.6 m s^{-1}
8. 4 s
9. 125 m, 2.5 m s^{-2}
10. 15 m
11.



12.



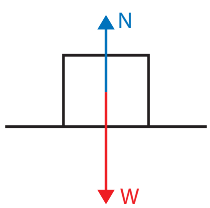
13. 78 m

14. **(a)** 30° **(b)** 17.3 m

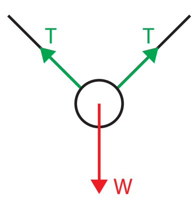
15. 5 m

16. 39.6 m

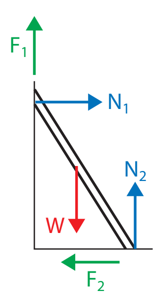
17. **(a)**



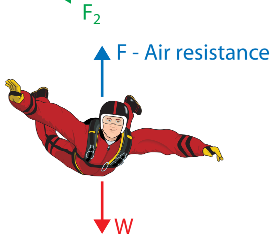
(b)



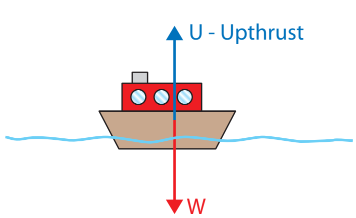
(c)



(d)



(e)



18. **(a)** 10 N **(b)** 5.8 N

19. **(a)** $F = T \sin 30°$ **(b)** $10 = T \cos 30°$
(c) 11.5 N **(d)** 5.8 N

20. **(a)** $F = 50 \sin 30°$ **(b)** $50 \cos 30° = N$
(c) F=25 N, N =43.3 N

21. **(a)** $600N = 2T \cos 80°$ **(b)** $T \sin 80° = T \sin 80°$
(c) 1728 N

22. **(a)** 4 N down slope **(b)** 4 N, 37.6°

23. **(a)** F1 = 8.49 N
(b) F2 = 17.3 N, F3 = 50 N

24. −3 N s

25. −4.0 N s

26. 6.7 m s^{-2}

27. 997.5 N

28. **(a)** 3.3 m s^{-2} **(b)** 33 N

29. 2 m s^{-2}

30. 682.5 N

31. **(a)** 40 N **(b)** 150 m s^{-2}

32. **(a)** force on gas = −force on rocket
(b) force on water = −force on boat
(c) force on body = −force on board
(d) Water exerts unbalanced force on ball, so ball exerts force on water; reading increases.

33. **(a)** 9 m s^{-1} **(b)** −1 m s^{-1} **(c)** 0.85 m s^{-1}

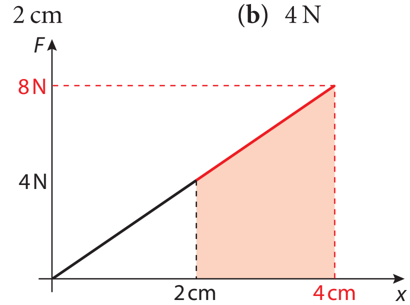
34. **(a)** 0.875 N s **(b)** 44 m s^{-1}

35. **(a)** −1300 J **(b)** Dog

36. 0 J

37. −300 J

38. **(a)** 2 cm **(b)** 4 N



(d) 4 J **(e)** 12 J

39. 1950 J

40. 32 m s^{-1}

41. **(a)** 12.5 J **(b)** 12.5 J **(c)** 1 cm

42. **(a)** 0.75 J **(b)** 0.75 m

43. **(a)** 0.2 m **(b)** 0.8 m

44. **(a)** 1.25×10^{-4} J
(b) 6.25×10^{-5} J
(c) 0.1 m s^{-1}

45. **(a)** −10.83 m s^{-1} **(b)** 521.3 J

46. Velocities swap

47. 800 W

48. 1000 W

49. 20 kW

50. 50%

51. 42 kJ

52. **(a)** 6.67 kW **(b)** 11.1 kW

53. 1389 N

54. 15.8 m s^{-1}

55. **(a)** 3.14 m **(b)** 1.57 s **(c)** 2π rad
(d) 4 rad s^{-1} **(e)** 8 m s^{-2} **(f)** 2 N

Practice questions

1. **(a)** **(i)** $18t$; 1

(ii) $s = \frac{1}{2} \times 4.5 \times 6^2 = 81$ m; 1

(iii) $v = at = 6 \times 4.5 = 27\ \mathrm{m\,s^{-1}}$; 1

(iv) $27(t - 6)$; 1

(b) idea of **(a)** **(i)** = **(a)** **(ii)** + **(a)** **(iv)**;

$18t = 81 + 27(t - 6)$

$t = 9.0$ s; 2

[6]

2. **(a)** statement that gravitational mass and inertial mass have the same numerical value;

understanding of what gravitational mass means;

e.g. "a quantity that determines the gravitational force on the object"

understanding of what inertial mass means;

e.g. "a quantity that determines the acceleration of the object" 3 max

(b) **(i)** the acceleration = gradient of first section of graph;

acceleration $= \dfrac{0.80}{0.50} = 1.6\ \mathrm{m\,s^{-2}}$; 2 max

Accept bald correct answer for full marks.

(ii) the total distance travelled by the lift = area under graph;

distance $= (11 \times 0.80) + (0.50 \times 0.80)$

$= 8.8 + 0.4 = 9.2$ m; 2 max

Accept bald correct answer for full marks.

(iii) the work done = P.E. gained

(= force × distance);

work done $= 2500 \times 9.2 = 23\,000$ J $= 23$ kJ; 2 max

Accept bald correct answer for full marks.

(iv) correct substitution into power

$= \dfrac{\text{work done}}{\text{time taken}}$

$= \dfrac{23000}{12}$;

$= 1916$ W

$= 1.9$ kW; 2 max

(v) correct substitution into efficiency

$= \dfrac{\text{power out}}{\text{power in}}$

$= \dfrac{1.9}{5.0}$

$= 0.38 = 38\ \%$; 2 max

(c) graphs should show curving or "shoulders" at the changes;

since acceleration must be finite / speed cannot change instantaneously / *OWTTE*;

2 max

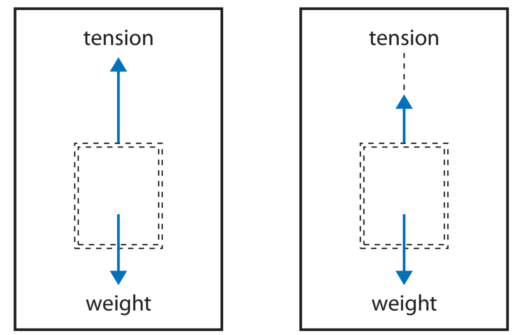
(d) *Mark part (i) and (ii) together.*

weight arrow the same in both diagrams;

magnitude of tension (size of arrow) **equal to** weight in **(i)**;

magnitude of tension (size of arrow) **less than** weight in **(ii)**;

(i) 0.50 to 11.50 s **(ii)** 11.50 to 12.00 s

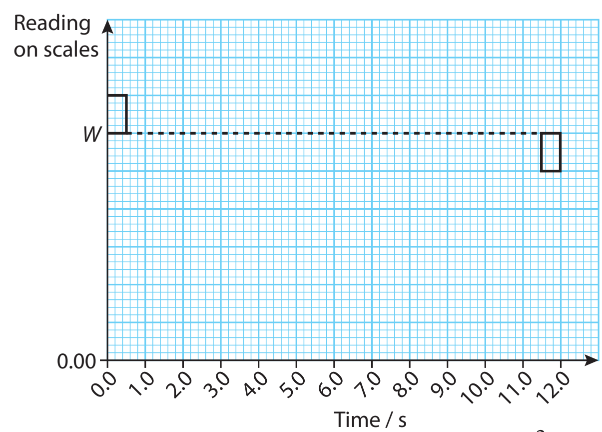


3 max

(e) a constant value **greater** than W from 0.00 to 0.50 s;

a constant value **equal** to W from 0.50 to 11.50 s;

a constant value **less** than W from 11.50 to 12.00 s;



3 max

(f) *[1] for each appropriate and valid point. Essentially [2] for journey up and [2] for journey down. Some explanation or justification is required for full marks e.g.*

the law of conservation of energy does apply to round trip;

energy is all dissipated into heat and sound;

on the way up, most electrical energy converted into gPE, initially

some electrical energy is converted into KE;

on the way down electrical energy does work "breaking" lift some

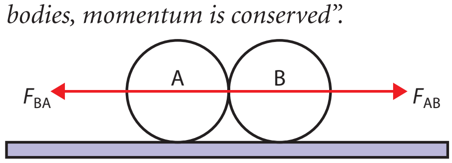
(not all) gPE is converted into KE; 4 max

Reject answers that imply that P.E. converts into KE as lift falls.

[25]

3. (a) (i) $h = \dfrac{v^2}{2g}$;
to give $h = 3.2$ m; 2
(ii) 0.80 s; 1
(b) time to go from top of cliff to the sea
$= 3.0 - 1.6 = 1.4$ s;
recognise to use $s = ut + \frac{1}{2}at^2$ with correct substitution,
$s = 8.0 \times 1.4 + 5.0 \times (1.4)^2$;
to give $s = 21$ m; 3
Answers might find the speed with which the stone hits the sea from $v = u + at$, (42 ms^{-1}) and then use $v^2 = u^2 + 2as$. [6]

4. (a) when two bodies A and B interact, the force that A exerts on B is equal and opposite to the force that B exerts on A;
or
when a force acts on a body an equal and opposite force acts on another body somewhere in the universe; 1 max
*Award **[0]** for "action and reaction are equal and opposite" unless they explain what is meant by the terms.*
(b) if the net external force acting on a system is zero;
then the total momentum of the system is constant (or in any one direction, is constant); 2
*To achieve **[2]** answers should mention forces and should show what is meant by conserved. Award **[1 max]** for a definition such as "for a system of colliding bodies, the momentum is constant" and **[0]** for "a system of colliding bodies, momentum is conserved".*
(c)



arrows of equal length;
acting through centre of spheres;
correct labelling consistent with correct direction; 3
(d) (i) *Ball B:*
change in momentum $= Mv_B$;
hence $F_{AB}\Delta t = Mv_B$; 2
(ii) *Ball A:*
change in momentum $= M(v_A - V)$;
hence from Newton 2, $F_{BA}\Delta t = M(vA - V)$; 2
(e) from Newton 3, $F_{AB} + F_{BA} = 0$, ***or*** $F_{AB} = -F_{BA}$;
therefore $-M(v_A - V) = Mv$;
therefore $MV = Mv_B + Mv_A$;
that is, momentum before equals momentum after collision such that the net change in momentum is zero (unchanged) / *OWTTE*; 4
Some statement is required to get the fourth mark i.e. an interpretation of the maths result.
(f) from conservation of momentum $V = v_B + v_A$;
from conservation of energy $V^2 = v_B{}^2 + v_A{}^2$;
if $v_A = 0$, then both these show that $v_B = V$;
or
from conservation of momentum $V = v_B + v_A$;
from conservation of energy $V^2 = v_B{}^2 + v_A{}^2$;
so, $V^2 = (v_B + v_A) = v_B{}^2 + v_A{}^2 + 2v_Av_B$
therefore v_A has to be zero; 3 max
Answers must show that effectively, the only way that both momentum and energy conservation can be satisfied is that ball A comes to rest and ball B moves off with speed V. [17]

5. (a) the direction of the car is changing;
hence the velocity of the car is changing;
or
since the direction of the car is changing;
a force must be acting on it, hence it is accelerating; 2 max
(b) (i) arrow pointing vertically downwards; 1
(ii) weight;
normal reaction; 2
(iii) loss in PE $= 0.05 \times 10 \times (0.8 - 0.35)$;
$=$ gain in KE $= \frac{1}{2}mv^2$;
to give $v = 3.0\ \mathrm{m\,s^{-1}}$;
or
use of $v = \sqrt{2gh}$ to give $v = 4.0\ \mathrm{m\,s^{-1}}$ at point B;
and then use of $v^2 - u^2 = 2gh$ with $v = 4.0\ \mathrm{m\,s^{-1}}$ and $h = 0.35$ m;
to get $u = 3.0\ \mathrm{m\,s^{-1}}$; 3 max
(iv) recognize that resultant force $= \dfrac{mv^2}{r}$;
$= \dfrac{(0.05 \times 9.0)}{0.175} = 2.6$ N;
$N = \dfrac{mv^2}{r} - mg$;
$= 2.6 - 0.5 = 2.1$ N 4
[12]

6. (a) mass $\times$ velocity; 1
(b) (i) momentum before $= 800 \times 5 = 4000$ N s;
momentum after $= 2\,000v$;
conservation of momentum gives
$v = 2.0\ \mathrm{m\,s^{-1}}$; 3

(ii) KE before $= 400 \times 25 = 10\,000$ J
KE after $= 1000 \times 4 = 4000$ J;
loss in KE $= 6000$ J; 2

(c) transformed/changed into;
heat (internal energy) (and sound); 2
Do not accept "deformation of trucks".

[8]

7. (a) ***Note**: for part* **(i)** *and* **(ii)** *the answers in brackets are those arrived at if 19.3 is used as the value for the height.*

(i) height raised $= 30 \sin 40 = 19$ m;
gain in PE $= mgh = 700 \times 19$
$= 1.3 \times 10^4$ J (1.4×10^4 J); 2

(ii) $48 \times 1.3 \times 10^4$ J
$= 6.2 \times 10^5$ J (6.7×10^5 J); 1

(iii) the people stand still / don't walk up the escalator / their average weight is 700 N / ignore any gain in KE of the people; 1 max

(b) (i) power required $= \dfrac{6.2 \times 10^5}{60}$
$= 10$ kW (11 kW);
$Eff = \dfrac{P_{out}}{P_{in}}, P_{in} = \dfrac{P_{out}}{Eff}$;
$P_{in} = 14$ kW (16 kW); 3

(ii) the escalator can in theory return to the ground under the action of gravity / *OWTTE*; 1

(c) power will be lost due to friction in the escalator / *OWTTE*; 1
The location of the friction must be given to obtain the mark.

[9]

Chapter 3

1. (a) 7.12×10^{-6} m^3 (b) 6.022×10^{23} molecules
(c) 1.2×10^{-29} m^3
2. 27 g
3. (a) 3.92×10^3 J (b) 3.92×10^3 J
4. 1.8×10^6 J
5. 420 kJ
6. (a) 3.6×10^6 J (b) 3.6×10^5 J °C^{-1}
(c) Some is lost to the outside.
7. 1.33×10^4 J
8. (a) 1kg (b) 3.36×10^5 J (c) 336 s
9. (a) 3×10^5 J (b) 686.7 °C
10. (a) 3×10^5 J (b) 2.25×10^5 J
(c) 51°C
11. (a) 80 kg (b) 1.34×10^7 J
12. 3.35×10^{11} J
13. 1.135×10^3 s
14. (a) 1.84×10^4 kg (b) 6.16×10^9 J
(c) 3.42×10^5 W (d) 342 W m^{-2}
15. (a) 292 kPa
16. (a) 6 kPa (b) 3 kPa
17. 312.5 cm^3
18. 400 kPa
19. (a) P↑ V↑ T↑ (b) P↓ T↓
(c) P↑ V↓ (d) V↑ T↑
20. 20 J
21. 37.5 J
22. (b) 600 K, 1200 K, 600 K (c) 25 J
(d) 50 J (e) −25 J
23. (a) 245 J (b) 50 J
(c) 320 J (d) 135 J
(e) 185 J
24. (a) (i) −1.25 J K^{-1}
(ii) 2 J K^{-1}
(b) 0.75 J K^{-1}

Practice questions

1. (a) (165, 0); 1

(b) *Look for these points:*
to change phase, the separation of the molecules must increase;
Some recognition that the ice is changing phase is needed.
so all the energy input goes to increasing the PE of the molecules;
Accept something like "breaking the molecular bonds".
KE of the molecules remains constant, hence temperature remains constant; 3
If KE mentioned but not temperature then assume they know that temperature is a measure of KE.

(c) (i) time for water to go from 0 to 15 °C = 30 s;
energy required $= ms\Delta\theta$
$= 0.25 \times 15 \times 4\,200 = 15\,750$ J;
power $= \dfrac{\text{energy}}{\text{time}} = 525 \text{ W} \approx 530 \text{ W}$; 3

(ii) ice takes 15 s to go from −15 °C to 0;
energy supplied $= 15 \times 530$ J;
sp ht $= \dfrac{(530 \times 15)}{(15 \times 0.25)} = 2\,100 \text{ J kg}^{-1} \text{ K}^{-1}$; 3

(iii) time to melt ice = 150 s;
$L = \dfrac{(150 \times 530)}{0.25} = 320 \text{ kJ kg}^{-1}$; 2

[12]

2. (a) more energetic molecules leave surface;
mean kinetic energy of molecules in liquid decreases;
and mean kinetic energy depends on temperature; 3
Award [2] if mean not mentioned.

(b) *e.g.* larger surface area;
increased draught;
higher temperature;
lower vapour pressure; 2 max
Award [1] if candidate merely identifies two factors.

(c) energy to be extracted $= 0.35 \times 4\,200 \times 25$;
$+0.35 \times 330\,000$;
$+0.35 \times 2\,100 \times 5$;
$= 156\,000$ J

time $= \frac{156\,000}{86} = 1800$ s; 4
Allow ECF if one term is incorrect or missing.
[9]

3. Increasing temperature increases speed of molecules. This results in increased change in momentum on colliding with smaller walls. Increased rate of change of momentum means greater force, and since area is constant, greater pressure. [3]

4. (a) (i) $120\ \text{J kg}^{-1}\,\text{K}^{-1}$ (ii) $5.4 \times 10^5\ \text{J kg}^{-1}$

5. (a) *[1] for each appropriate and valid point e.g.*
thermal energy is the KE of the component particles of an object;
thus measured in joules;
the temperature of an object is a measure of how hot something is
(it can be used to work out the direction of the natural flow of thermal
energy between two objects in thermal contact)/measure of the average
KE of molecules;
it is measured on a defined scale (Celsius, Kelvin *etc.*); 4 max

(b) (i) correct substitution: energy
= power × time;
$= 1200\ \text{W} \times (30 \times 60)$ s;
$= 2.2 \times 10$ J 2 max

(ii) use of $E = mc\Delta\theta$,
to get $\Delta\theta = 2.2 \times 10^6/(4\,200 \times 70)$ K;
$= 7.5$ K; 3 max

(c) *[1] naming each process up to **[3 max]**.*
convection;
conduction;
radiation;
*[1] for an appropriate (matching) piece of information/outline for each process up to **[3 max]**.*
e.g. convection is the transfer of thermal energy via bulk movement of a gas due to a change of density;
conduction is transfer of thermal energy via intermolecular collisions;
radiation is the transfer of thermal energy via electromagnetic waves
(IR part of the electromagnetic spectrum in this situation)/*OWTTE*; 6 max

(d) (i) *[1] for each valid and relevant point e.g.*
in evaporation the faster moving molecules escape;
this means the average KE of the sample left has fallen;
a fall in average KE is the same as a fall in temperature; 3 max

(ii) energy lost by evaporation
$= 50\% \times 2.2 \times 10^6$ J;
$= 1.1 \times 10^6$ J;
correct substitution into $E = ml$
to give mass lost
$= 1.1 \times 10^6\ \text{J}/(2.26 \times 10^6)\ \text{J kg}^{-1}$
$= 0.487$ kg
$= 487$ g; 3 max

(iii) *[1] for any valid and relevant factors **[2 max]***
e.g. area of skin exposed;
presence or absence of wind;
temperature of air;
humidity of air *etc.*;
*[1] for appropriate and matching explanations **[2 max]***
e.g. increased area means greater total evaporation rate;
presence of wind means greater total evaporation rate;
evaporation rate depends on temperature difference;
increased humidity decreases total evaporation rate *etc.*; 4 max
[25]

6. D

7. C

8. A

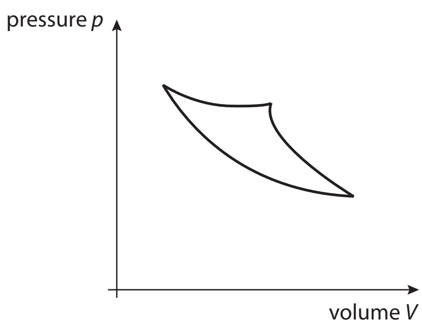
9. A

10. C

11. (a) statement (implication) that work done is associated with area
within the rectangle;
Do not award mark for just "area" without reference.
calculation of $2 \times 10^5 \times 8 = 1.6 \times 10^6$ J; 2 max

(b) thermal energy from hot reservoir
$= 1.8 \times 10^6 + 1.6 \times 10^6$ J
$= 3.4 \times 10^6$ J;
efficiency
= work done / thermal energy from hot reservoir
$= 1.6 \times 10^6 / 3.4 \times 10^6 = 47\ \%$; 2 max
[0] for $\frac{1.6 \times 10^6}{1.8 \times 10^6} = 89\ \%$.

(c) closed cycle of rough approximate shape;
quality of diagram (adiabatic "steeper" than isothermal *etc.*); 2 max



(d) **(i)** adiabatic (expansion and contraction);
isothermal (expansion and contraction); 2 max

(ii) correct "sense" of adiabatic followed by and isothermal *etc.*;
e.g. adiabatic (expansion) then
isothermal (contraction) then
adiabatic (contraction) then
isothermal (expansion) then
correct identification of adiabatic as the steeper curve when compared with isothermal; 2 max

[10]

12. **(a)** isothermal: takes place at constant temperature;
adiabatic: no energy exchange between gas and surroundings; 2

(b) **(i)** neither; 1

(ii) $\Delta W = P\Delta V = 1.2 \times 10^5 \times 0.05$
$= 6.0 \times 10^3$ J; 1

(iii) recognize to use $\Delta Q = \Delta U + \Delta W$;
to give $\Delta U = 2.0 \times 10^3$ J; 2

[6]

13. **(a)** **(i)** on – gas is compressed 1 max
Correct answer and correct explanation.

(ii) ejected – pressure remains constant, volume reduced so temperature must go down 1 max
Correct answer and correct explanation.

(b) work done $= p\Delta V$;
$= -1.0 \times 10^5 \times 0.4 = -0.40 \times 10^5$ J (40 kJ); 2 max
Sign should be consistent with **(a) (i)** *above – work "by" and + work here would get zero for* **(a) (i)** *but [2] marks here.*

(c) area enclosed;
$\approx 0.6\ (\pm 0.2) \times 10^5$ J (60 kJ $\pm$ 20 kJ); 2 max

(d) efficiency = work out/heat in;
$= \frac{60}{120} = 50\ \%\ (\pm 17\ \%)$; 2 max

[8]

14. **(a)** pV constant for isothermal / adiabatic always steeper; **hence** AB; 2

(b) area between lines AB and AC shaded; 1

(c) area is 150 ($\pm$15) small squares;
(allow ecf from (b))
work done $= 1.5 \times 1 \times 10^{-3} \times 1 \times 10^5$;
$= 150$ J; 3
*For any reasonable approximate area outside the range 150 (±15) squares award **[2 max]** for the calculation of energy from the area.*

(d) no thermal energy enters or leaves / $\Delta Q = 0$;
so work done seen as increase in internal energy;
hence temperature rises; 3
*Award **[0]** for a mere quote of the 1st law.*

[9]

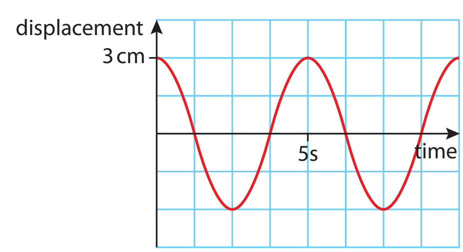
Chapter 4

1. **(b)**

2. **(a)** 1.67 Hz **(b)** 10.5 Hz

3. 1.1 cm down

4.



5. 1.67s

6. -0.5 m s^{-1}

7. **(a)** 2.5 m s^{-1} **(b)** 3.16 m s^{-2}

8. 0.37 m s^{-1}

9. 0.318 m

10. 29°

11. 15°

12. 2×10^8 m s^{-1}

13. $\frac{\pi}{10}, \frac{\pi}{2}, \frac{3\pi}{4}$

14. **(a)** 182.6 m s^{-1} **(b)** 143.8 Hz

15. 165 Hz

16. **(a)** 32.22 cm **(b)** 2

17. **(a)** oooeeee **(b)** 1100 Hz **(c)** 892 Hz

18. 321.8 m s^{-1}

19. 318.2 Hz
20. 0.06 c
21. 9.2×10^7 Hz
22. 4.38 nm
23. 0.11 m
24. 88 μm
25. 7×10^7 m
26. 14.6 m
27. No, 0.0001 rad < 0.00014 rad
28. 56°
29. Reflected light cannot pass through.
30. $I_o/8$
31. **(a)** 5 W m^{-2} **(b)** 0 W m^{-2} **(c)** 1.25 W m^{-2}

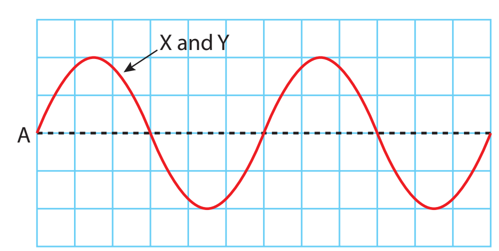
Practice questions

1. **(a)** longitudinal; 1
(b) **(i)** wavelength = 0.5 m; 1
(ii) amplitude = 0.5 mm; 1
(iii) correct substitution into speed = frequency × wavelength;
to give $v = 660 \times 0.5 = 330$ m s^{-1}; 2 max
[5]

2. **(a)** wavefront parallel to D; 1
(b) **(i)** frequency is constant;
since, $v = f\lambda$, $v \propto \lambda$
(ii) wavelength larger in medium I, **hence** higher speed in medium I; 3
Allow solution based on angles marked on the diagram or speed of wavefronts.
(c) **(i)** velocity/displacement/direction in (+) and (−) directions;
idea of periodicity; 2
(ii) period = 3.0 ms;
frequency $= \frac{1}{T} = 330$ Hz; 2
(iii) *Accept any one of the following.*
at time $t = 0$ / 1.5 ms / 3.0 ms / 4.5 ms *etc.*; 1 max
(iv) area of half-loop = 140 squares ± 10 / mean $v = 4.0$ m s^{-1} ± 0.2;
$= 140 \times 0.4 \times 0.1 \times 10^{-3}$ or $4.0 \times 1.5 \times 10^{-3}$;
$= 5.6 \times 10^{-3}$ m or 6.0×10^{-3} m; 2 max
(v) (twice) the amplitude; 1
Allow distance moved in 1.5 ms.

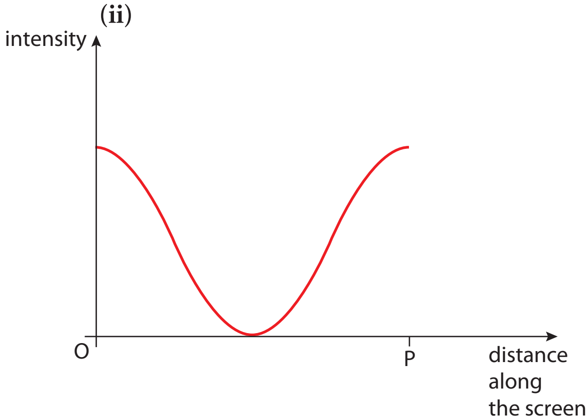
3. **(a)** **(i)** Distance travelled by a fixed point on the wave per unit time
(ii) velocity has direction; but light travels in all directions; 2
(b) **(i)** *longitudinal*: displacement along;
transverse: displacement normal to;
(ii) direction of transfer of wave energy/ propagation, **not** motion; 3
Award **[0]** *for left/right and up/down for longitudinal/transverse.*
(c) **(i)** $\left(\frac{700}{75}\right) = 9.3$ km s^{-1}; (±0.1) 1
(ii) $\left(\frac{700}{120}\right) = 5.8$ km s^{-1}; (±0.1) 1
Award **[1 max]** *if the answers to* **(i)** *and* **(ii)** *are given in reversed order.*
(d) **(i)** P shown as the earlier (left hand) pulse; 1
(ii) Laboratory L_3; 1
(iii) *e.g.* pulses arrive sooner;
smaller S-P interval;
larger amplitude of pulses; 3
Allow any feasible piece of evidence, award **[1]** *for each up to* **[3 max]**.
(iv) *distance from* $L_1 = 1060$ km; (±20)
distance from $L_2 = 650$ km; (±20)
distance from $L_3 = 420$ km; (±20)
Accept 3 significant digits in all three estimates.
some explanation of working; 4
(v) position marked, consistent with answers to **(iv)**;
to the right of line L_2L_3, closer to L_3; 1 max

4. **(a)** the net **displacement** of the medium/particles (through which waves travel);
is equal to the sum of individual **displacements** (produced by each wave); 2 max
Award a good understanding **[2 max]** *and a reasonable one* **[1 max]**.
(b) *Wave X and wave Y should be identical.*



correct phase for wave X;
correct phase for wave Y;
amplitudes the same for each wave;
amplitude for each wave is two divisions; 4 max
(c) **(i)** the phase difference between light leaving S_1 and S_2 is constant; 1
Do not penalize the candidate if they state "has the same phase".
(ii) to produce **sufficient** diffraction;
for the beams to overlap;
OWTTE; 2 max
(d) **(i)** path difference between S_1 and S_2 is an integral number of wavelengths; 1
Accept "waves arrive at P in phase".

(ii)



maximum at O and P;
general shape with minimum about half way between O and P; 2 max

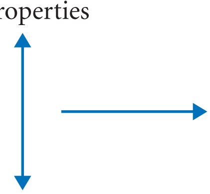
(e) fringe spacing $= 2.5 \times 10^{-4}$ m;

$\lambda = \dfrac{(2.5 \times 10^{-4} \times 3.00 \times 10^{-3})}{1.50} = 5.0 \times 10^{-7}$ m; 2

[14]

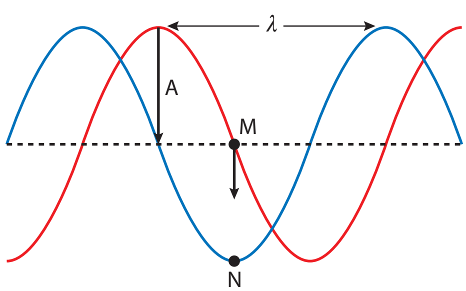
5. Wave properties

(a) **(i)** 1



(ii) 1

(b)



(i) downwards; 1

(ii) correct marking of A; 1

(iii) correct marking of λ; 1

(iv) +ve sine curve;
correct position of N; 2
Watch for ecf from **(i)**.

(c) **(i)** $f = \dfrac{v}{\lambda} =$ to give 2.0 Hz; 1

(ii) $T = 0.5$ s;

$s = \dfrac{vT}{4} = 1.25\ (1.3)$ cm;

or

in $\dfrac{T}{4}$ wave moves forward $\frac{1}{4}\lambda$;

$= \frac{5}{4} = 1.25\ (1.3)$ cm; 2 max

(d) *Principle of superposition*:
when two or more waves overlap, the resultant displacement at any point;
is the sum of the displacements due to each wave separately / *OWTTE*;
*Award [**2 max**] for an answer that shows a clear understanding of the principle, [**1**] for a reasonable understanding and [**0**] for a weak answer.*
Explanation:

suitable diagram;
when two +ve pulses (or two wave crests) overlap, they reinforce / OWTTE; 4
Any situation where resultant displacement looks as though it is the sum of the individual displacements. Mark the description of the principle and the description of constructive interference together.

(e) **(i)** $S_2X = n\lambda$;
where $n = 0, 1, 2$; *(Accept "n is an integer")* 2

(ii) $\sin\theta \approx \theta$;
therefore $\theta = \dfrac{S_2X}{d}$; 2

(iii) $\phi = \dfrac{y_n}{D}$; 1

Award the small angle approximation mark anywhere in (i) or (ii).

(f) **(i)** $\theta = \dfrac{S_2X}{d}$ *so* $\lambda = \dfrac{d\theta}{n}$;
substitute to get $\lambda = 4.73 \times 10^{-7}$ m; 2

(ii) θ and ϕ are small;
therefore $\dfrac{\lambda}{d} = \dfrac{y}{D}$;
so $y = D\dfrac{\lambda}{d} = 0.510$ mm; 3

[24]

6. **(a)** circular wavefronts originating from four successive source positions;
bunching of wavefronts in front, spreading out at back;
approximately, correct spacing of wavefronts in front, and behind source; 3

(b) f waves in distance $(V - v)$;
apparent wavelength $= \dfrac{(V - v)}{f}$;
apparent frequency $= \dfrac{f \times V}{(V - v)}$; 3
*Allow any other valid and correct approach or statement of formula. Award [**0**] for quote of formula with no working shown.*

(c) $\lambda' = \lambda\frac{(V - v)}{V}$;

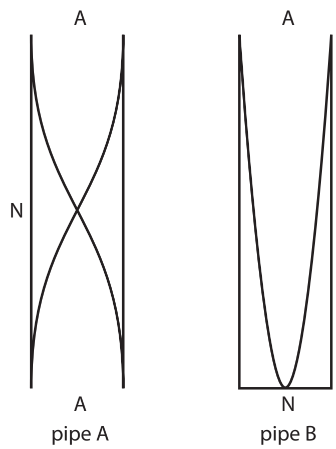
$599.996 = \frac{600 \times (3 \times 10^8 - v)}{(3 \times 10^8)}$;

$v = 2000\ \text{m s}^{-1}$; 3

Allow alternative version for red-shift.

[9]

7. (a)



(i) correct wave shape for pipe A;
correct wave shape for pipe B; 2

(ii) correct marking of A and N for pipe A;
correct marking of A and N for Pipe B; 2

(b) (i) for pipe A, $\lambda = 2L$, where L is length of the pipe;
$c = f\lambda$ to give $L = \frac{c}{2f}$;
substitute to get $L = 0.317$ m; 3

(ii) for 32 Hz, the open pipe will have a length of about 5 m;
whereas the closed pipe will have half this length, so will not take up as much space as the open pipe / *OWTTE*; 2
*The argument does not have to be quantitative. Award **[1]** for recognition that low frequencies mean longer pipes and **[1]** for the same frequency, closed pipes will be half the length of open pipes. The fact they need less space can be implicit.* [9]

8. (a) (i) diffraction at the lens; 1

(ii) circular patch – bright;
circular bright ring/darkness between patch and ring; 2

(b) (i) $\alpha = \frac{4.0 \times 10^{-6}}{17 \times 10^{-3}}$;
$= 2.4 \times 10^{-4}$ rad; 2

(ii) $1.22\frac{\lambda}{d} = 2.4 \times 10^{-4}$ therefore
$d = 1.22 \times 550 \times 10^{-9}/2.4 \times 10^{-4}$;
$d = 2.8$ mm; 2
*Award **[2]** even if factor 1.22 is missing.*

[6]

Chapter 5

1. 3 kΩ
2. 0.3 V
3. 0.02 A
4. 100 kΩ, 100 kΩ, 25 kΩ
5. 1 Ω
6. 11.5 V
7. (a) 500 J (b) 3×10^4 J
8. 0.03125 W
9. 0.5 W
10. (a) 450 kJ (b) 37.5 kW
(c) 125 A (d) No other losses
11. (a) 0.45 A (b) 20 J
12. (a) 4.5 A (b) 1.8×10^7 J
13. 16/3 Ω
14. 8 Ω
15. 28 Ω
16. 16/7 Ω
17. 5 V, 0.5 A
18. 3 V, 3 A
19. 6 V, 1.5 A
20. 6 V, 3 A
21. 714 Ω

Practice questions

1. (i) use of emf = energy / charge;
$= \frac{(8.1 \times 10^3)}{(5.8 \times 10^3)}$
$= 1.4$ V; 2
*Award **[0]** for formula $E = \frac{F}{Q}$ seen or implied even if answer is numerically correct.*

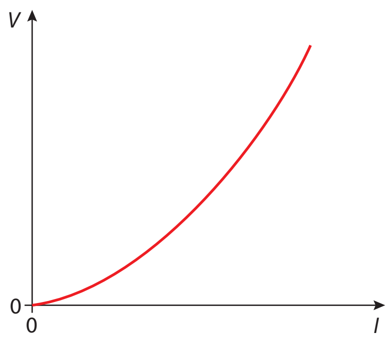
(ii) p.d. across internal resistance = 0.2 V;
current $= \frac{1.2}{6} = 0.2$ A;
resistance $r = \left(\frac{0.2}{1.2}\right) \times 6.0$;
$= 1.0\ \Omega$;
or
total resistance $= \frac{1.4}{0.2} = 7.0\ \Omega$;
internal resistance $= 7 - 6 = 1.0\ \Omega$; 3
Accept any other valid route.

(iii) idea of use of ratio of resistances;
energy transfer $= 6/7 \times 8.1 \times 10^3$
$= 6.9(4) \times 10$ J; 2
Accept any other valid route.

(iv) charge carriers/electrons have kinetic energy / are moving;
these carriers collide with the lattice / lattice ions; *(do not allow friction)*

causing increased (amplitude of) vibrations;
this increase seen as a temperature rise;
i.e. a transfer to thermal energy; 5
Allow any other relevant and correct statements.

2. **(a)**



1

Any reasonable curve in the right direction.

(b) **(i)** from the value of V/I at any point on the curve; 1
Do not accept just "from V/I".

(ii) non-ohmic because the resistance (V/I at each point) is not constant/ *OWTTE*; 1

(c) **(i)** 50 Ω; 1

(ii) recognize that the voltage must divide in the ratio 3 : 1;
to give R = 150 Ω; 2 max
Or answer could be solved via the current.

[6]

3. **(a)** **(i)** when connected to a 3 V supply, the lamp will be at normal brightness;
and energy is produced in the filament at the rate of 0.60 W;
Look for the idea that 3 V is the operating voltage and the idea of energy transformation.
or
when connected to a 3 V supply, the lamp will be at normal brightness;
and the resistance of the filament is 15 Ω/ the current in the filament is 0.20 A; 2 max

(ii) $I = \frac{P}{V}$;
to give $I = 0.20$ A; 2

(b) **(i)** at maximum value, the supply voltage divides between the resistance of the variable resistor, internal resistance and the resistance of the filament;
i.e. response must show the idea of the voltage dividing between the various resistances in the circuit. Do not penalise if responses do not mention internal resistance here.
at zero resistance, the supply voltage is now divided between the filament resistance and the internal resistance of the supply; 2

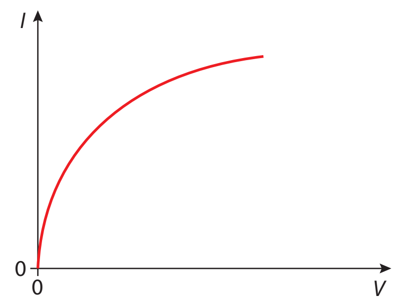
(ii) when resistance of variable resistor is zero, e.m.f. $= Ir + V_{lamp}$;
$3.0 = 0.2r + 2.6$;
to give $r = 2.0$ Ω; 3

(c) **(i)** 3.3 Ω; 1

(ii) 13 Ω; 1

(d) at the higher p.d., greater current and therefore hotter;
the resistance of a metal increases with increasing temperature; *OWTTE*; 2 max

(e)



correct approximate shape (*i.e.* showing decreasing gradient with increasing V); 1

(f) parallel resistance of lamp and YZ is calculated from $\frac{1}{R} = \frac{1}{4} + \frac{1}{12}$;
to give $R = 3.0$ Ω;
3.0 V therefore divides between 3.0 Ω and 12.0 Ω;
to give p.d. across the lamp = 0.60 V;
Give relevant credit if answers go via the current i.e.
calculation of total resistance = 15.0 Ω;
total current = 0.20 A;
current in lamp = 0.15 A; 4

[18]

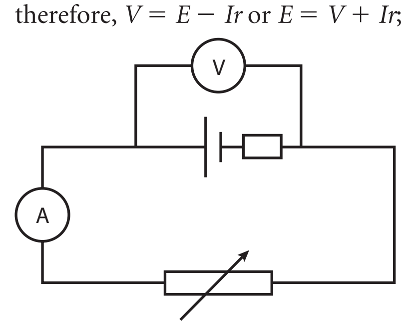
4. **(a)** **(i)** EI; 1

(ii) I^2r; 1

(iii) VI; 1

(b) (from the conservation of energy), $EI = I^2r + VI$;
therefore, $V = E - Ir$ or $E = V + Ir$; 2

(c)



correct position of voltmeter;
correct position of ammeter;
correct position of variable resistor; 3

(d) (i) $E = V$ when $I = 0$;
so $E = 1.5\ V$; 2

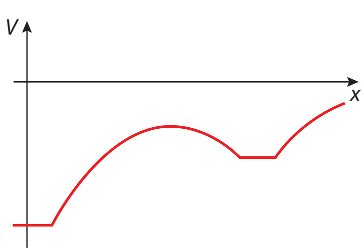
(ii) recognize this is when $V = 0$;
intercept on the x-axis $= 1.3\ (\pm\ 0.1)$ A; 2

(iii) r is the slope of the graph;
sensible choice of triangle, at least half the line as hypotenuse;
$= \frac{0.7}{0.6}$;
$= 1.2\ (\pm\ 0.1)\ \Omega$;
or
when $V = 0$, $E = Ir$;
$r = \frac{E}{I}$
$= \frac{1.5}{1.3}$
$= 1.2\ \Omega$; 3

(e) $R = 1.2\ \Omega$;
$I = \frac{1.5}{1.2 + 1.2} = 0.63$ A;
$P = I^2R = (0.63)^2 \times 1.2 = 0.48$ W / 0.47 W; 3
[18]

Chapter 6

1. 1.62 m s^{-2}
2. 24.7 m s^{-2}
3. 7.34 m s^{-2}
4. 6.69×10^{-8} N kg^{-1}
5. 0 N kg^{-1}
6. 30 J kg^{-1}
7. 90 J
8. 240 J
9. 0
10. 0
11. (a) 1.55 MJ kg^{-1}
(b) 3.1×10^9 J
(c)

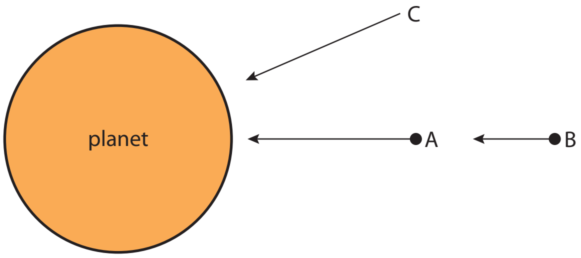


13. Hydrogen would escape.
14. 3 km
15. 2.74 km s^{-1}
16. Graph of T^2 vs r^3
17. 4.2×10^7 m
18. 1.5 hr
19. (a) -5.9×10^{10} J (b) -1.2×10^{11} J
(c) 6.1×10^{10} J
20. (a) 1.8×10^6 N C^{-1} (b) 4.5×10^5 N C^{-1}
(c) 0.045 N
21. (a) 1×10^{-7} N (b) 1×10^{-5} m s^{-2}
22. 2 V
23. 5 V
24. 15 J
25. 4 J
26. -8 J
27. (a) It accelerates downwards.
(b) 12 J
28. 2.25×10^6 V
29. 1.13×10^6 V
30. (a) Q1 +ve (b) towards Q2
31. F
32. (a) 20 V (b) 10 V (c) 0 V
33. (a) 40J (b) -20J (c) 0J
34. 50 Vm^{-1}
35. -3nC
36. (a) -10 eV (b) -50 eV (c) 20 eV
37. (a) 2×10^{-5} N (b) East
38. (a) 5×10^{-6} N (b) West
39. (a) 2×10^{-4} V (b) 1×10^{-4} A
(c) 2×10^{-8} J (d) 2×10^{-8} J
(e) 2 m (f) 1×10^{-8} N
40. (a) 1×10^{-6} T m^{-2}
(b) 0.25×10^{-6} T m^{-2} s^{-1}
(c) 0.25 μV
41. (a) 1.5×10^{-5} T m^{-2} (b) 1.3×10^{-5} T m^{-2}
(c) 0.67 μV
42. 156 V
43. 18A
44. (a) (i) 100π rad s^{-1}
(ii) 3.9 V
(iii) 2.8 V
(b) 1.4 V
45. 48.4 Ω
46. (a) 10 (b) 2 W
(c) 9.1 mA (d) 0 A
47. (a) 5×10^3 A (b) 200 MW
(c) 40% (d) 300 MW
(e) 300 MW (f) 1.36 MA

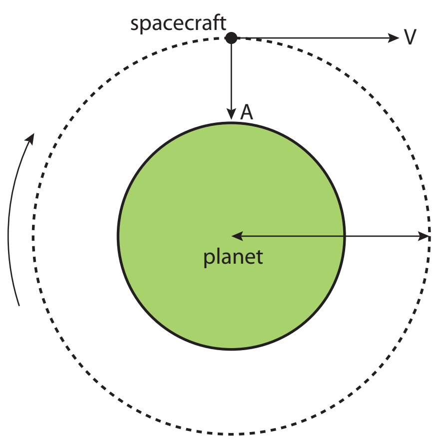
Practice questions

1. (a) $g = \frac{F}{m}$
F is a the gravitational force;
exerted on/experienced by a small/point/infinitesimal mass m; 2

(b) *Award [1] for each correct arrow. The one at B points in the same direction as that at A and is shorter. The one at C has the same length as t hat at A and points towards the centre of the planet.* 2



(c)



(i) velocity is tangent to path; 1

(ii) acceleration is normal to velocity towards centre; 1

(d) for realizing that $g = a$;

$a = \frac{v^2}{r} = \frac{(6.5 \times 10^3)^2}{7.5 \times 10^6} \text{N kg}^{-1}$;

$g = 5.6 \text{ N kg}^{-1}$; 3

[9]

2. (a) attractive force is proportional to the product of the point masses;

and inversely proportional to the square of the separation; 2

Award [1] if the response is not clear that they are point masses or if the force is attractive.

Award [0] for quoting the formula from data booklet without any further explanation.

(b) use of $g = \frac{GM}{r^2}$;

appropriate substitution:

$g = \frac{6.67 \times 10^{-11} \times 6.0 \times 10^{24}}{(6.4 \times 10^6)^2} = 9.77$

$\approx 10 \text{ N kg}^{-1}$; 2

(c) **(i)** point marked on Earth's surface that is nearest to Moon;

since force of attraction from Moon greatest; 2

Accept other sensible comment.

(ii) same point as above;

Accept point on directly opposite side of Earth.

explanation of why the resultant field is a minimum at this point

e.g. forces from Earth and Moon are in opposite directions; 2

(iii) each relevant point;

e.g. Earth is rotating;

Moon orbits the Earth *etc.*;

position of Sun also affects resultant field *etc.*; 2 max

[10]

3. (a) the force exerted per unit mass;

on a point/small mass; 2

(b) **(i)** use of $g = \frac{F}{m}$ and $F = G\frac{Mm}{R^2}$;

combine to get $g = G\frac{M}{R^2}$; 2

(ii) $M = \frac{gR^2}{G}$;

substitute to get $M = 1.9 \times 10^{27}$; 2

[6]

4. (a) The work done per unit mass;

in bringing a (small test) mass from infinity to the point; 2

Idea of ratio crucial for first mark.

(b) **(i)** $g = \frac{GM}{r_1^{\,2}} - \frac{Gm}{r_2^{\,2}}$;

$0 = \frac{M}{0.8^2} - \frac{m}{0.2^2}$;

$\frac{M}{m} = 16$; 3

(ii) $K = m\Delta V$;

$K = 1500 \times (4.6 - 0.20) \times 10^7$;

$K = 6.6 \times 10^{10}$ J; 3

*Award **[2 max]** if attempted use of ΔV but value used is wrong and **[1 max]** if an individual potential value rather a difference is used.)* [8]

5. (a) **(i)** $V_{\text{surface}} = -6.3\ (\pm\ 0.3) \times 10^7 \text{ J kg}^{-1}$ 1 max

(ii) V_{h} is at $R = 42 \times 10^6$ m;

$= -1.0\ (\pm 0.2) \times 10^7 \text{ J kg}^{-1}$; 2 max

Watch for $R = 3.6 \times 10^7$ m being used. If so award [1] and use ECF.

(b) $\Delta V = 5.3\ (\pm\ 0.5) \times 10^7 \text{ J kg}^{-1}$;

Energy $= m\Delta V$;

$= 5.3\ (\pm\ 0.5) \times 10^{11}$ J; 3 max

Award [2] if they calculate the PE of the satellite (10^{11} J).

(c) *Any **two** of the following [1] each.*
the satellite has to be given a horizontal velocity (or has to have KE) to go into orbit;
rockets motors lifting rocket not 100 % efficient;
air resistance in initial stages of launch; 2 max
[8]

6. **(a)** force exerted per unit mass;
on a small / point mass; 2

(b) from the law of gravitation, the field strength
$\frac{F}{m} = G\frac{M}{R^2}$; $= g_0$ to give $GM = g_0R^2$; 2
N.B. *To achieve full marks, candidates need to state that* $\frac{F}{m} = g_0$.

(c) downwards; (*accept 90° to B field or down the wire*) 1

(d) $F = Bev\cos\theta$; 1

(e) work done in moving an electron the length of the wire is
$W = FL = BevL\cos\theta$;
emf = work done per unit charge;
therefore, $E = BLv\cos\theta$;
or
electric field $= \frac{F}{e} = Bv\cos\theta$;
emf E = electric field $\times L$;
to give $E = BLv\cos\theta$; 3
*Award [**2 max**] if flux linkage argument is used.*

(f) $F = G\frac{Mm}{r^2} = \frac{mv^2}{r}$;
such that $v^2 = \frac{GM}{r} = \frac{g_0R^2}{r}$;
$v^2 = \frac{10 \times (6.4)^2 \times 10^{12}}{6.7 \times 10^6}$
to give $v = 7.8 \times 10^3$ ms^{-1}; 3

(g) $L = \frac{E}{Bv\cos\theta}$;
$= \frac{10^3}{6.3 \times 10^{-6} \times 7.8 \times 10^3 \times 0.93}$
$= 2.2 \times 10^4$ m; 2
[14]

7. **(a)** **(i)** <u>emf</u> (induced) proportional to;
rate of change /cutting of (magnetic) flux (linkage); 2

(ii) magnetic field / flux <u>through coil</u> will change as the current changes; 1

(b) **(i)** sinusoidal and in phase with current; 1

(ii) sinusoidal and same frequency;
with 90° phase difference to candidate's graph for φ; 2

(iii) e.m.f. is reduced;
because B is smaller; 2
*Award [**0**] for "e.m.f. is reduced" if argument fallacious.*

(c) advantage: no direct contact with cable required;
disadvantage: distance to wire must be fixed; 2
[10]

8. A [1]

9. C [1]

Chapter 7

1. **(a)** 9.6×10^{-20} J **(b)** 7.1×10^{14} Hz
(c) 3.7×10^{-19} J **(d)** 5.7×10^{14} Hz

2. **(a)** 8.6 eV **(b)** 4.3 eV
(c) 4.3 V **(d)** 1.0×10^{15} Hz

3. No

4. 1.5×10^{15} Hz

5. 10

6. 13.06 eV, 3.17×10^{15} Hz

7. 0.31 eV

8. 13.6 eV, 3.29×10^{15} Hz

9. **(a)** 100 eV **(b)** 1.6×10^{-17} J
(c) 1.2×10^{-10} m

10. 4.4×10^{-38} m

11. **(a)** 10 000 m s^{-1} **(b)** 57.5 U

12. 3.2 cm

13. **(a)** 1.9×10^7 m s^{-1} **(b)** 4.9×10^{-15} m

14. **(a)** 17p, 18n **(b)** 28p, 30n
(c) 82p, 122n

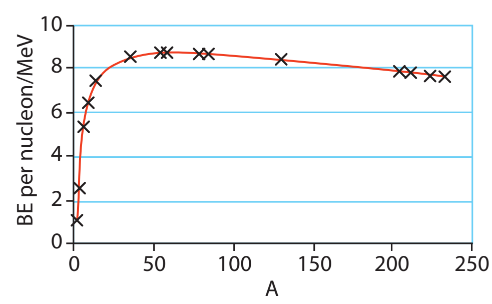
15. 4.16×10^{-18} C, 9×10^{-26} kg

16. $^{235}_{92}$U

17. 92 protons, 146 neutrons

18. **(a)** 92p, 141n **(b)** 234.9405u
(c) 1.901u **(d)** 1771 MeV
(e) 7.60 MeV

19.



20. 21p, 24n

21. 55p, 82n

22. **(a)** α **(b)** α **(c)** β

23. **(a)** 5.24 MeV **(b)** 4.67 MeV
(c) 0.39 MeV

24. 12.5 g

25. 12.5 s^{-1}

26. 24 000 years

27. 7.54Bq

28. 28.6 yr
29. **(a)** 10^{11} nuclei **(b)** 6.5×10^8 yr
30. 912 yr
31. **(a)** 1.66×10^8 s **(b)** 4.17×10^{-9} s^{-1}
(c) 1.0×10^{22} **(d)** 4.17×10^{13} s^{-1}
(e) 1.2×10^{-12} g
32. **(a)** $^{18}_{9}$F, $^{1}_{1}$H **(b)** $^{123}_{53}$I
(c) $^{17}_{7}$N, $^{18}_{8}$O **(d)** $^{30}_{15}$P
33. **(a)** $^{-}1.017375$ U **(b)** 948 MeV
34. **(a)** 3.27 MeV **(b)** 4.03 MeV
(c) 18.3 MeV
35. 10n, 133.9 MeV
36. 135.7 MeV

Practice questions

1. **(a)** *Deduct [1] for each error or omission, stop at zero.* 2 max

Property	Effect on rate of decay		
	increase	decrease	stays the same
temperature of sample			✓
pressure on sample			✓
amount of sample	✓		

(b) **(i)** $^4_2$He$^4_2\alpha$;
$^{222}_{86}$Rn; 2
(ii) mass defect $= 5.2 \times 10^{-3}$ u;
energy $= mc^2$
$= 5.2 \times 10^{-3} \times 1.661 \times 10^{-27} \times 9.00 \times 10^{16}$
$= 7.77 \times 10^{-13}$ J or 4.86 MeV; 3 max
(c) **(i)** (linear) momentum must be conserved;
momentum before reaction is zero;
so equal and opposite after (to maintain zero total); 3
(ii) $0 = m_\alpha v_\alpha + m_{\text{Rn}} v_{\text{Rn}}$;
$\frac{v_\alpha}{v_{\text{Rn}}} = -\left(\frac{m_{\text{Rn}}}{m_\alpha}\right)$
$= -\frac{222}{4} = -55.5$; 3
Ignore absence of minus sign.
(iii) kinetic energy of α-particle $= \frac{1}{2} m_\alpha v_\alpha{}^2$;
kinetic energy of radon nucleus
$= \frac{1}{2}\left(\frac{222}{4}\right) m_\alpha \left(\frac{v_\alpha}{55.5}\right)^2$;
this is 1 / 55.5 of kinetic energy of the α-particle; 3 max
Accept alternative approaches up to [3 max].

(d) *e.g.* (γ-ray) photon energy or radiation; 1
(e) **(i)** two (light) nuclei;
combine to form a more massive nucleus;
with the release of energy / with greater total binding energy; 3
(ii) high temperature means high kinetic energy for nuclei;
so can overcome (electrostatic) repulsion (between nuclei);
to come close together / collide;
high pressure so that there are many nuclei (per unit volume);
so that chance of two nuclei coming close together is greater; 5
[25]

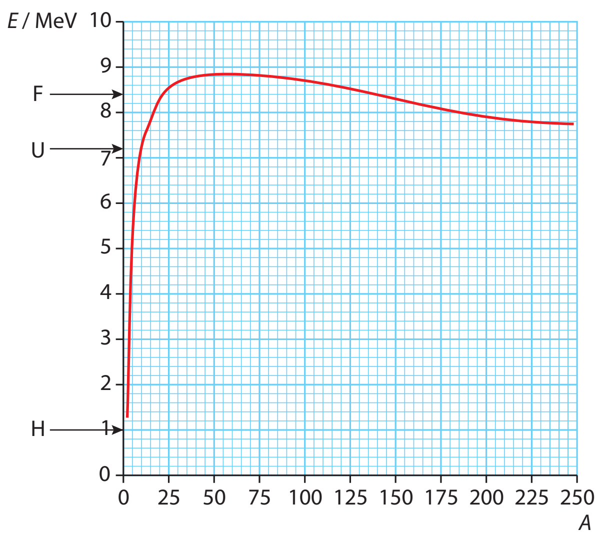
2. **(a)** **(i)** *fission*:
nucleus splits;
into two parts of similar mass;
radioactive decay:
nucleus emits;
a particle of small mass and/or a photon; 4
(ii) $^{235}_{92}$U + $^{1}_{0}$n;
$\rightarrow$ $^{90}_{38}$Sr + $^{142}_{54}$Xe + 4^{1}_{0}n; 2
Allow ecf for RHS if LHS is incorrect.
(iii) mass number unchanged;
atomic number increases by +1; 2
(b) **(i)** use of $E_k = \frac{p^2}{2m}$ / equivalent;
correct conversion of MeV to joules (1.63×10^{-11} J);
correct conversion of mass to kilogram (1.50×10^{-25} kg);
momentum $= 2.2 \times 10^{-18}$ N s; 4
(ii) total momentum after fission must be zero;
must consider momentum of neutrons (and protons); 2
(iv) xenon not opposite to strontium but deviation <30°);
arrow shorter / longer; 2
(c) **(i)** energy $= 0.25 \times 198 \times 1.6 \times 10^{-13}$;
$= 7.9 \times 10^{-12}$ J; 2
(ii) use of $\Delta Q = mc\Delta Q$;
energy $= 0.25 \times 4\,200 \times 80$;
$= 8.4 \times 10^4$ J; 3
(iii) number of fissions $= \frac{(8.4 \times 10^4)}{(7.9 \times 10^{-12})}$;
mass $= 1.1 \times 10^{16} \times 3.9 \times 10^{-25}$;
$= 4.1 \times 10^{-9}$ kg; 4
[25]

3. **(a)** **(i)** a proton or a neutron; 1

Both needed to receive [1].

(ii) the difference between the mass of the nucleus and the sum of the masses of its individual nucleons/the energy required to separate a nucleus into its component nucleons/*OWTTE*; 1

(b)



Don't expect precision for any of these.

(i) F: between 8 and 9; 1

(ii) H: between 1 and 2; 1

(iii) U: between 7 and 8; 1

(c) general overall shape;

max at F = 56, end point U; 2

(d) mass of nucleons = $(2 \times 1.00728) + 1.00867$

$= 3.02323$ u;

mass difference = 0.0072 u = 6.7 MeV;

binding energy per nucleon = 6.7/3

= 2.2 MeV; 3

(e) **(i)** fusion; 1

(ii) from the position on the graph, the energy required to assemble two nuclei of ${}^{2}_{1}$H is greater than that to assemble one nucleus of ${}^{3}_{2}$He;

hence if two nuclei of ${}^{2}_{1}$H combine to form one nucleus of ${}^{3}_{2}$He energy must be released/*OWTTE*; 2

[13]

4. **(a)** all particles have a wavelength associated with them / *OWTTE*;

the de Broglie hypothesis gives the associated wavelength as $\lambda = \frac{h}{p}$;

where h is the Planck constant and p is the momentum of the particle; 3

If answers just quote the formula from the data book then award [1] for showing at least they recognize which formula relates to the hypothesis.

(b) **(i)** KE $= Ve = 850 \times 1.6 \times 10^{-19}$ J

$= 1.4 \times 10^{-16}$ J; 1

(ii) use $E = \frac{p^2}{2m}$ to get $p = \sqrt{2mE}$;

substitute

$p = \sqrt{2 \times 9.1 \times 10^{-31} \times 1.4 \times 10^{-16}}$

$= 1.6 \times 10^{-23}$ N s; 2

(iii) $\lambda = \frac{h}{p}$;

substitute

$\lambda = \frac{6.6 \times 10^{-34}}{1.6 \times 10^{-23}} = 4.1 \times 10^{-11}$ m; 2

[8]

5. **(a)** *Mark the both processes, 1 and 2, together. Award [1] any two of the following.*

collisions with (external) particles;

heating the gas to a high temperature;

absorption of photons; 2 max

(b) **(i)** $E = \frac{hc}{\lambda}$

$E = \frac{6.63 \times 10^{-34} \times 3 \times 10^{8}}{658 \times 10^{-9}}$;

$E = \frac{3.02 \times 10^{-19}}{1.6 \times 10^{-19}}$ eV;

$= 1.89$ eV 2

(ii) electrons absorb photons (of energy 1.89eV) to make a transition from $n = 2$ to $n = 3$;

on de-excitation, photons of energy 1.89eV, *i.e.* wavelength 658 nm are emitted;

in **all directions**, however, and not just along the initial direction,

hence intensity is reduced; 3

(iii) (the Schrödinger model unlike Bohr's) does not have well defined orbits for the electrons / does not treat the electron as a localized particle / assigns to an electron a probability wave;

predicts the relative intensities of various spectral lines; 2

[9]

6. **(a)** *aspect*:

electrons will not be emitted unless the frequency of light exceeds a certain minimum value / electrons are emitted almost instantaneously with the light falling on the surface even if light is of very low intensity / the energy of the electrons emitted is not affected by the intensity of light falling on the surface; 1

corresponding explanation:
light consists of photons whose energy is hf hence no electrons are emitted unless hf is larger than the energy needed to escape the metal / an electron is emitted as soon as it absorbs a photon. If the photon has sufficient energy no delay is required / the intensity of light plays no role in the energy of the electron only the frequency of light does; 1

(b) **(i)** the threshold frequency is found from the frequency axis intercept;
to be $3.8(\pm 0.2) \times 10^{14}$ Hz; 2

(ii) a value of the Planck constant is obtained from the slope;
to be $6.5(\pm 0.2) \times 10^{-34}$ J s; 2
Award ***[0]*** *for "bald" answer of 6.63×10^{-34} J s.*

(iii) the work function of the surface is found from the intercept with the vertical axis;
to be $1.5(\pm 0.1)$ eV; 2

(c) straight–line parallel to the first;
intersecting the frequency axis at 8.0×10^{14} Hz; 2
[10]

7. **(a)** $^{40}_{19}K \rightarrow ^{40}_{18}Ar + \beta^+(e^+) + \nu$
β^+ / e^+; 2 max

(b) 8.2×10^{-6} g; 1

(c) **(i)** $\lambda = \ln 2 / T_{\frac{1}{2}}$;
$$\lambda = \frac{0.69}{1.3 \times 10^9} = 5.3 \times 10^{-10}\ \text{year}^{-1};$$ 2

(ii) from $N = N_0 e^{-\lambda t}$ $\quad t = \frac{1}{\lambda}\ln\left(\frac{N_0}{N}\right)$;
$= 1.9 \times 10^9 \times \ln(6.8) = 3.6 \times 10^9$ years ;
or
$$\frac{1.2}{8.2} = \left(\frac{1}{2}\right)^n$$
$n = 2.77$;
age $= 2.77 \times 1.3 \times 10^9 = 3.6 \times 10^9$ years; 2
[7]

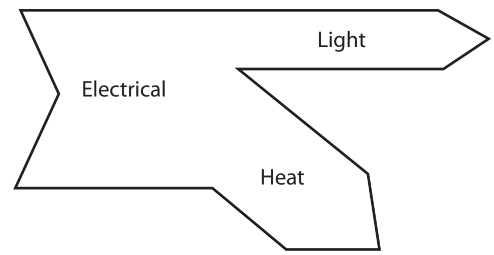
8. **(a)** $qvB = m\frac{v^2}{r}$;
hence $r = \frac{mv}{Bq}$; 2

(b) $$\frac{16.5}{15} = \frac{\frac{m_{16.5}v}{Bq}}{\frac{m_{15}v}{Bq}} = \frac{m_{16.5}}{m_{15}};$$
hence $\frac{16.5}{15} = \frac{m_{16.5}}{20} \Rightarrow m_{16.5} = 22u$; 2

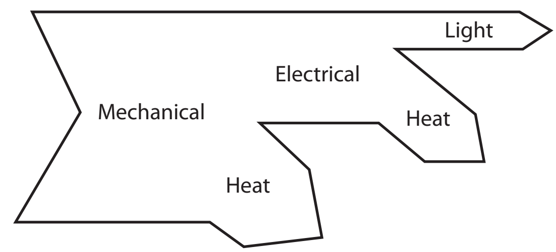
(c) atoms on 15 cm path: 10 protons and 10 neutrons;
atoms on 16.5 cm path: 10 protons and 12 neutrons; 2
[6]

Chapter 8

1.



2.



3. **(a)** 80 km **(b)** 1.44×10^8 J
(c) 1.92×10^8 J **(d)** 4.2 kg
(e) 4.7 litres **(f)** $0.06\ \text{l km}^{-1}$

4. **(a)** 8.64×10^{13} J **(b)** 2.16×10^{14} J
(c) 6.65×10^6 kg **(d)** 67 truck loads

5. **(a)** $^{238}_{92}U \rightarrow ^{0}_{1}n \rightarrow ^{239}_{92}U$
(b) $^{239}_{92}Np \rightarrow ^{239}_{93}Np + \beta^- + \bar{\nu}$
(c) $^{239}_{93}U + ^{239}_{94}Pu + \beta^- + \bar{\nu}$

6. **(a)** $^{142}_{56}Ba \rightarrow ^{142}_{57}La + \beta^- + \bar{\nu}$
(b) 9 years

7. **(a)** 7
(b) $^{239}_{94}Pu \rightarrow ^{136}_{54}Xe + ^{96}_{40}Zr + 7n^0$
(c) 164 MeV **(d)** 239 g
(e) 2.5×10^{24} atoms **(f)** 4.13×10^{26} MeV
(g) 6.6×10^{13} J

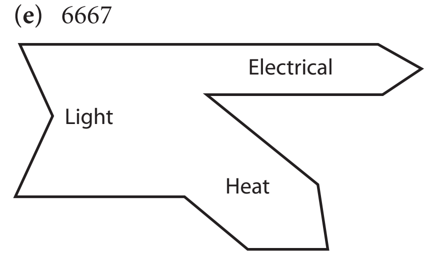
8. 2.7×10^{12} J

9. **(a)** 3.6×10^{10} J **(b)** 13g

10. **(a)** 4000 W **(b)** 2000 J
(c) 28.6 °C

11. **(a)** 0.015 W **(b)** 0.03 A
(c) 5 V **(d)** 0.3 A
(e) 6667

12.



13. **(a)** 1.5×10^{14} kg **(b)** 1.6×10^{17} J
(c) 1.9×10^9 J **(d)** 1.7×10^6 kg

14. **(a)** 3.2 kW **(b)** 64 kg

15. **(a)** 5.5 MW **(b)** 1.1 MW
(c) 3.7 MW

16. **(a)** $10\ \text{m s}^{-1}$ **(b)** 49 kW **(c)** 98 MW
17. **(a)** $9552\ \text{W m}^{-2}$ **(b)** $51.3\ \text{W m}^{-2}$
18. 3.3×10^{-19} J (2eV)
19. 4.8×10^{-7} m
20. $7.35 \times 10^{7}\ \text{W m}^{-2}$
21. 4.52×10^{26} W

Practice questions

1. **(a)** (natural process of) production takes thousands/millions of years;
fossil fuels used much faster than being produced/*OWTTE*; 2

(b) *Any two sensible suggestions e.g.*
storage of radioactive waste;
increased cost;
risk of radioactive contamination *etc.*; 2 max
To achieve full marks the differences must be distinct.

[4]

2. **(a)** solar panel: solar energy → thermal energy (heat);
solar cell: solar energy → electrical energy; 2

(b) **(i)** input power required = 730 W (± 5) W;
area $= \frac{730}{800} = 0.91\ \text{m}^2$; 2

(ii) power extracted = 165 W (±20 W);
efficiency $= \frac{\text{(power out)}}{\text{power in}}$
or
$\frac{165}{500}$; *(allow ecf)*
= 33%; 3

[7]

3. **(a)** idea of thermal energy → mechanical energy/KE → electrical energy;
idea of where or how this takes place; 2
e.g. in turbines or coil rotated in a magnetic field etc.

(b) *Mark the answers for the two energy sources together.*
both non renewable;
appropriate justification for both; 2
e.g. in both cases a resource is being used and isn't being replaced/OWTTE.

(c) **(i)** to slow down fast moving neutrons;
so as to increase chances of neutron capture by another uranium nucleus/*OWTTE*; 2

(ii) to absorb neutrons;
so as to control rate of reaction/*OWTTE*;2

(d) any appropriate advantage that coal fired power station does not have;
e.g. does not release CO_2/SO_2 into atmosphere/OWTTE.
appropriate discussion relating to advantage;
e.g. so global warming/acid rain effects reduced. 2
*Allow argument that 1 kg of uranium "fuel" releases more energy w.r.t. 1 kg of coal. Award **[0]** for imprecise statements that are not clear e.g. bald "nuclear power stations pollute less".*

[10]

4. **(a)** power $= \frac{\text{energy}}{\text{time}} = \frac{120 \times 10^{12}}{60 \times 60 \times 24 \times 365}$;
$= 3.8 \times 10^{6}\ W$;
therefore, for one turbine = 0.19 MW; 3

(b) using $p = \frac{1}{2}\rho A v^3$, $A = \frac{2p}{\rho v^3}$;
therefore, $A = \frac{2 \times 1.9 \times 10^5}{1.2 \times 9.0^3} = 4.3 \times 10^2\ \text{m}^2$;
use $A = \pi r^2$ to give $r = 12$ m; 3

(c) the wind speed varies over the year / not all the wind energy will be transferred into mechanical power / energy loss due to friction in the turbine / energy loss in converting to electrical energy / density of air varies with temperature;
Do not accept something like "turbines are not 100% efficient".

(d) take up so much room;
that not possible to produce enough energy to meet a country's requirements;
noisy;
and this could have an effect on local fauna;
OTTWE; 2 max
*Award **[1]** for statement of disadvantage and **[1]** for some justification of statement.*

[9]

5. **(a)** **(i)** fission; 1 max
(ii) kinetic energy; 1 max

(b) the two neutrons can cause fission in two more uranium nuclei producing four neutrons so producing eight *etc.*; *OTTWE*;
1 max

(c) **(i)** the fuel rods contain a lot more U-238 than U-235;
neutron capture is more likely in U-238 than U-235 with high energy neutrons;
but if the neutrons are slowed they are more likely to produce fission in U-235 than neutron capture in U-238; 3 max
The argument is a little tricky so be generous. The candidate needs to know about there being two isotopes present in the fuel and something about the

dependence of the fission and capture in the two isotopes on neutron energy.

(ii) control the rate at which the reactions take place;
by absorbing neutrons; 2 max

(d) *Look for* ***four*** *of the following main points and award* ***[1]*** *each.*
energy lost by the slowing of the neutrons and fission elements heats the pile;
this heat extracted by the molten sodium/ pressurized water/other suitable substance;
which is pumped to a heat exchanger;
water is pumped through the heat exchanger and turned to steam;
the steam drives a turbine;
which is used to rotate coils (or magnets) placed in a magnetic field (or close to coils)
which produces electrical energy; 4 max
Alternatively, award ***[4]*** *for a good answer,* ***[2]*** *for a fair answer and* ***[1]*** *for a weak answer.*

[12]

6. **(a)** **(i)** kinetic energy of the fission products/ neutrons/photons; 1
Do not accept thermal energy or heat.

(ii) if mass of uranium is too small too many neutrons escape;
without causing fission in uranium/ reactions cannot be sustained; 2

(iii) the moderator (and the fuel rods);

(iv) mass of uranium atom
$= 235 \times 1.661 \times 10^{-27} = 3.90 \times 10^{-25}$ kg
or $\frac{0.235}{6.02 \times 10^{23}}$ kg;
mass of uranium per second
$= 3.90 \times 10^{-25} \times 8 \times 10^{19}$
$= 3.12 \times 10^{-5}\ \text{kg s}^{-1}$;
mass of uranium per year
$= 3.12 \times 10^{-5} \times 365 \times 24 \times 60 \times 60$
$= 984 \approx 9.8 \times 10^{2}\ \text{kg yr}^{-1}$; 3

(b) **(i)** 800 MW; 1

(ii) 200 MW; 1

(iii) 1 600 MW; 1

(iv) the second law of thermodynamics; 1

(v) $\eta = \frac{800}{2400} = 33\%$; 1
Answer does not need to be expressed as a percentage.

(vi) $0.33 = 1 - \frac{300}{T_H}$;
$T_H = 450$ K; 2

[14]

7. **(a)** **(i)** $P \propto v^3$;
$= 15 \times 2^3 = 120$ kW; 2

(ii) the wind speed will not be reduced to zero after impact with blades;
power will be less because of frictional losses/turbulence; 2
The place where fractional losses take place must be identified.

(b) wind power is renewable;
while fossil fuels are finite; 2
Award ***[0]*** *for statement that wind generators do not cause pollution.*

[6]

Chapter 9

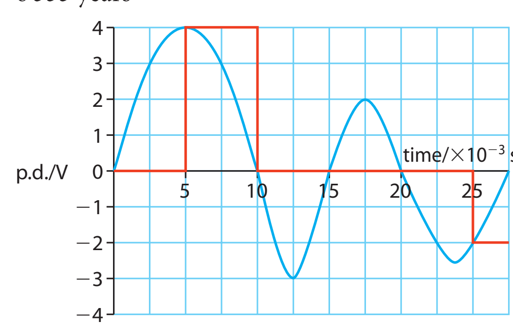
1. **(a)** 00000111
(b) 00001110
(c) 00010000
(d) 01000011
(e) 01111101

2. **(a)** 7 **(b)** 36
(c) 51 **(d)** 63

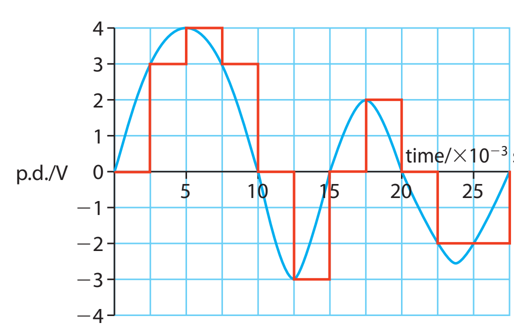
3. 01000010 01000001 01000100

4. 6 333 years

5.



6.



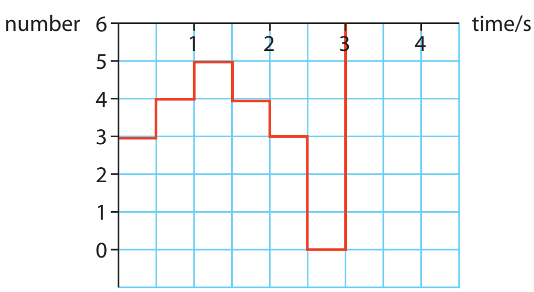
7. 100 Hz

8.

Time	No.	Binary
0.0	6	0110
0.5	10	1010
1.0	11	1011
1.5	9	1001
2.0	6	0110
2.5	2	0010
3.0	1	0001
3.5	2	0010
4.0	6	0110

9.

Time	No.	Binary
0.0	3	011
0.5	4	100
1.0	5	101
1.5	4	100
2.0	3	011
2.5	0	000



10. **(a)** 3×10^9 bits **(b)** 6×10^9 bits
(c) 753 MB

11. **(a)** 3.1×10^9 bits **(b)** 6.2×10^9 bits
(c) 781 MB

12. 0.33 V

13. 100 μC

14. 8×10^{-18} C 8×10^{-11} C

15. 4.24 mm 1.8×10^5 pixels

16. 4.24×10^{-4}

17. **(a)** 2×10^3 photons **(b)** 1.4×10^3 electrons

18. 0.2s

19. 1.7×10^5 pixels

Practice questions

1. A
2. C
3. D
4. C

5. **(a)** Magnification of camera = 3/6 000 = 5×10^{-4}
Separation of berries on image
$= 20 \times 5 \times 10^{-4} = 1 \times 10^{-2}$ mm
9 Mpixels means 3 000 × 3 000 so separation of pixels = 3/3 000 = 10^{-3} mm.
The image of each berry will appear on a different pixel so will be resolved.
(b) **(i)** 80% of photons liberate electrons so $0.8 \times 5\,000 = 4\,000$ electrons will be liberated in the pixel.
(ii) Charge of 4 000 electrons
$= 4\,000 \times 1.6 \times 10^{-19}$ C
(iii) $V = Q/C = \dfrac{6.4 \times 10^{-16}}{45 \times 10^{-12}} = 1.42 \times 10^{-5}$ V

6. **(a)** **(i)** 16 bits means that each voltage level is converted into a 16-bit digital number. 44 000 Hz means that the signal is sampled every 1/40 000 s.
(ii) In 5 minutes 5 × 60 × 44 000 readings are taken. 16 bits are used for each reading so a total of 5 × 60 × 44 000 × 16 = 2.11×10^7 bits are stored.

7. **(a)**

Time/ms	Voltage level	Binary
0	3	011
2	1	001
4	3	011
6	5	101
8	6	110
10	5	101
12	3	011

(b) Can be processed with a computer. Less susceptible to noise ***or*** can be compressed.

Chapter 10

2. 4.2 ly
3. 8 min 20 s
4. 1.5×10^5 yr
5. **(a)** 1.36×10^3 W m^{-2}
(b) 3.2×10^{-10} W m^{-2}
6. **(a)** 1.2×10^{-7} W m^{-2}
(b) 8.1×10^{-9} W m^{-2}
7. 5.6×10^3 ly
8. 4.2×10^{30} W
9. 1.6×10^8 W m^{-2}
10. **(a)** White dwarf **(b)** Red giant
(c) Main sequence

11. 7.5 hr
12. 3.1×10^{16}m
13. 0.2 arcsec
14. 6.7 pc
15. **(a)** 7250 K
(b) 3.84×10^{26} W
(c) 826 ly
16. **(a)** 2.69×10^{30} W
(b) 1730 ly
17. 4
18. 4.0×10^{-9} W m^{-2}
19. 6
20. 4×10^{9}
21. **(a)** 4.79×10^{-6} pc
(b) 8 pc
(c) 129 pc
(d) 131 pc
22. 3.17×10^{7} m s^{-1}
23. 3.25×10^{7} m s^{-1}
24. 2.1 Mpc
25. 1440 km s^{-1}

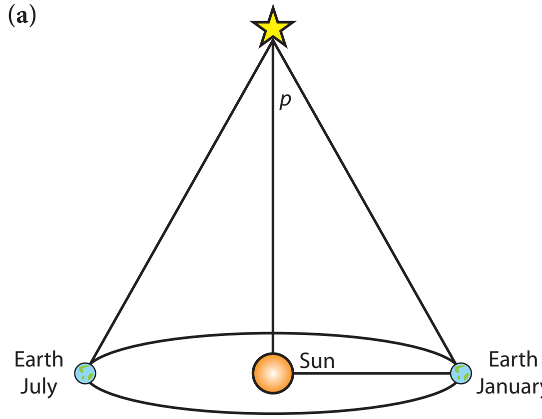
Practice questions

1. **(a)** **(i)** spectral class; 1
Accept colour sequence.
(ii) absolute magnitude; 1
(b) 4

Star	Type of star
A	Main sequence
B	Super Red Giant
C	White Dwarf
D	Main sequence

Award [1] for each correct name.
(c) B more luminous than A;
and has lower temperature than A;
so from the Stefan–Boltzmann law;
B has greater area (radius); 3 max
(d) use of $L = 4\pi bd^2$;
from the H-R diagram $L_B = 10^6\ L_{Sun}$;
therefore $\dfrac{L_B}{L_{sum}} = 10^6 = \dfrac{7.0 \times 10^{-8} \times d_B^{\,2}}{1.4 \times 10^3}$;
to give $d_B = 1.4 \times 10^8$ *AU* ($\approx$ 700 pc); 4
No mark is awarded for the conversion from AU to pc.
(e) at this distance the parallax angle is too small to be measured accurately;
OWTTE; 1 max
Do not accept "it's too far away"
[14]

2. **(a)**



Mark the definition of p and description of its measurement along with the diagram.
Essentially diagram should:
show p;
position of Sun;
position of Earth;
then
definition of $p = \dfrac{\text{(distance of Earth from Sun)}}{\text{(distance of start from Sun)}}$;
diagram should show Earth positions separated by about six months; then description should mention that angle of sight is measured at these two positions such that the difference between these two angles is equal to $2p$; 6 max
*Award **[6 max]** for a clear description and diagram, **[3]** for an average and **[1]** for some rudimentary idea. Mark diagram and description together.*
(b) $d = \dfrac{1}{p} = \dfrac{1}{0.549} = 1.82$ pc;
$= 1.82 \times 3.26 = 5.94$ ly; 2 max
(c) **(i)** the radiant power from a star;
that is incident per m^2 of the Earth's surface;
Alternatively, define from $b = \dfrac{L}{4\pi d^2}$ *but terms must be defined to obtain the mark.*
definition of L;
definition of d; 2 max
(ii) $L = 4\pi d^2 b$;
therefore, $\dfrac{L_B}{L_S} = \dfrac{d_B^{\,2} b_B}{d_s^{\,2} b_s}$;
$d_S = 1$AU, $d_B = 3.8 \times 10^5$ AU;
therefore, $\dfrac{L_B}{L_S} = (3.8)^2 \times 10^{10} \times 2.6 \times 10^{-14}$
$= 3.8 \times 10^{-3}$; 4 max
Allow any answer between (3.0 and 4.0) $\times 10^{-3}$.
(d) **(i)** temperature too low for it to be a white dwarf; 1 max

(ii) luminosity too low for it to be a red giant; 1 max

[16]

3. (a) (i) luminosity is the total power radiated by a star/source; 1

Do not accept $L = \sigma AT^4$.

(ii) apparent brightness is the power from a star received by an observer on Earth per unit area of the observer's instrument of observation; 1

Accept $b = \frac{L}{4\pi d^2}$ if L and d are defined.

(b) the surface area/size of the star changes periodically (due to interactions of matter and radiation in the stellar atmosphere); 1

(c) (i) at two days the radius is larger / point A; because then the luminosity is higher and so the area is larger; 2

Award [0] if no explanation is provided.

(ii) *Award [1] for each relevant and appropriate comment to the process of using Cepheid variables up to [**3 max**] e.g.*

Cepheid variables show a relationship between period and luminosity;

hence measuring the period gives the luminosity and hence the distance (through $b = \frac{L}{4\pi d^2}$);

distances to galaxies are then measured if the Cepheid can be ascertained to be within a specific galaxy; 3

Marks can be back credited from answer (d) (ii).

(d) (i) $b = \frac{L}{4\pi d^2} \Rightarrow 1.25 \times 10^{-10} = \frac{7.2 \times 10^{29}}{4\pi d^2}$;

$d = \sqrt{\frac{7.2 \times 10^{29}}{4\pi \times 1.25 \times 10^{-10}}}$;

$d = 2.14(\pm 0.2) \times 10^{19}$ m; 3

(ii) *Award [1] for each relevant and appropriate comment to the phrase "standard candles" up to [**2 max**] e.g.*

the phrase *standard candle* means having a source (of light) with known luminosity;

measuring the period of a Cepheid allows its luminosity to be estimated / other stars in the same galaxy can be compared to this known luminosity; 2

[13]

4. (a) apparent magnitude is a measure of (comparative) brightness as seen from Earth (with 1 being brightest and 6 being dimmest);

absolute magnitude is the apparent magnitude that the star would have if it were a fixed distance from the Earth of 10 parsecs; 2

(b) yes plus reason; 1

***Note**: an explanation must be provided. Award [0] for bald "yes" without an attempt at a reason. e.g. since apparent magnitude low (less than one) therefore one of the brightest stars.*

(c) (i) distance away $= \frac{3.39 \times 10^{17}}{9.46 \times 10^{15}} = 35.8$ ly $= 11.0$ pc; 1

(ii) since this is less than 100 pc;

the star is close enough for stellar parallax; 2

Award [1] for a bald answer. Also allow ECF if conversion of units is muddled.

(iii) *Award [1] each relevant piece of experimental description up to [**4 max**].*

e.g. position of star compared with other star positions;

at different times of the year;

the maximum angular variation from the mean p is recorded;

the distance (in parsecs) can be calculated using geometry $d = \frac{1}{p}$ if p is in arcseconds;

***Note**: watch for ECF. If the response has suggested one of the other techniques in (ii) then award full marks for appropriate descriptions.*

example:

spectroscopic parallax: light from star analysed (relative amplitudes of the absorption spectrum lines);

to give indication of stellar class;

HR diagram used to estimate the luminosity;

distance away calculated from apparent brightness;

Cepheid variables: these stars' brightness vary over time;

the time period of the variation is related to their luminosity;

thus measurements of the time period of one star can be used to calculate its luminosity;

its distance away is calculated from maximum apparent brightness; 4 max

(d) spectral type / K / *OWTTE*;

thus at low end of temperature scale: OBAFGKM / Sun is G / *OWTTE*; 2

(e) (i) correct substitution into $L = \sigma AT^4$;
to get $A = \dfrac{3.8 \times 10^{28}}{(5.67 \times 10^{-8} \times 4000^4)}$
$= 2.62 \times 10^{21}\ \mathrm{m}^2$; 2

(ii) use of $4\pi r^2 = 2.62 \times 10^{21}\ \mathrm{m}^2$;
to get $r = 1.44 \times 10^{10}$ m (= 0.10 AU); 2

(iii) use of $\lambda_{max} = \dfrac{2.90 \times 10^{-3}}{4000}$;
$= 725$ nm ≈ 730 nm; 2

(f) red giant;
since it's big and it's red / *OWTTE*; 2
[20]

5. **(a)** if less than ρ_0, Universe will expand for evermore;
if greater than ρ_0, Universe will expand;
and then contract; 3

(b) **(i)** substitution to give
$\rho_0 = 1.3 \times 10^{-26}\ \mathrm{kg\ m^{-3}}$; 1

(ii) number density $= \dfrac{(1.3 \times 10^{-26})}{(1.66 \times 10^{-27})}$, about 7 or 8 $\mathrm{m^{-3}}$; 1
***Note**: unit is m^{-3}.*
[5]

6. **(a)** **(i)** R shown amongst scattered points in upper right of diagram
W shown in lower region below main sequence, about centrally; 1

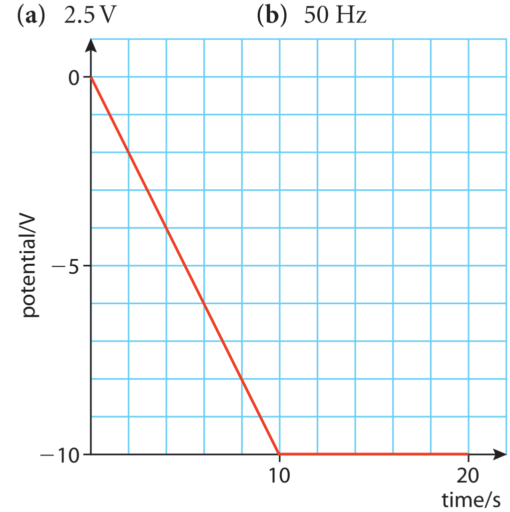
(ii) S shown on main sequence, about $\frac{1}{3}$ way up; 1
Allow the position of S anywhere between $\frac{1}{4}$ and $\frac{1}{2}$ the way up.

(iii) path shown to region of red giant;
then continuing to region of white dwarf; 2

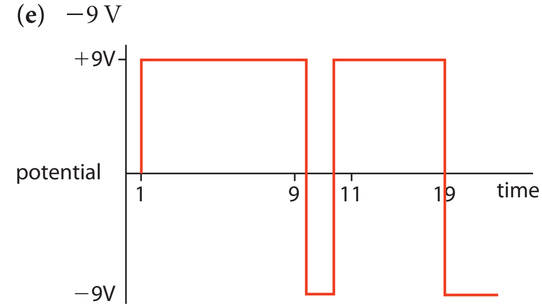
(b) (when forming a red giant) the star is expanding;
more power but over a much larger area, so cooler; 2
[6]

Chapter 11

1. 515 kHz, 485 kHz
2. **(a)** 1 Hz **(b)** 0.09 Hz **(c)** 0.91 Hz, 1.09 Hz
3. **(a)** 0.1 MHz **(b)** 0.001 **(c)** 240 kHz
4. **(a)** 0.33 MHz, 0.15 MHz **(b)** 0.05 MHz
5. **(a)** 001100 **(b)** 011010 **(c)** 100001
6. **(a)** 6 **(b)** 25 **(c)** 30
7. **(a)** 0 **(b)** 1 **(c)** 1
8. 101 101 011 001 000 101 010 010
9. 101 011 000 010
101 001 101 010
10. 5 3 0 2
5 1 5 2
11. 4, 6, 7, 6, 4, 1, 2, 4, 6, 4
12. 0100, 0110, 0111, 0110, 0100
13. 4.22×10^8 bits, 52.8MB
14. 94.5 min, 15.8 hours
15. **(a)** 1×10^{-6} s **(b)** 16 MHz
16. 5kHz
17. **(a)** 38.8° **(b)** 51.2° **(c)** 41.8° **(d)** yes **(e)** 62.2μm
18. **(a)** 10 dB **(b)** 7 dB **(c)** 20 dB
19. **(a)** 10 dB **(b)** 0.1 mW
20. No, attenuation too high
21. 8000 km
22. 25 000 km $\mathrm{h^{-1}}$
23. **(a)** 1600 km
24. **(a)** 5 V **(b)** 9 V **(c)** $^{-}$9 V
25. **(a)** 11 **(b)** 5.5 V **(c)** 0.5 mA **(d)** 5 V
26. **(a)** 5 **(b)** $^{-}$5 V **(c)** 1 mA **(d)** 1 mA **(e)** 1 V
27. **(a)** 2.5 V **(b)** 50 Hz
28.



29. 0.5 kΩ
30. **(a)** Reverse diode **(b)** 139.6 Ω **(c)** 7 mA **(d)** 0.976 V **(e)** −9 V
31.



Practice questions

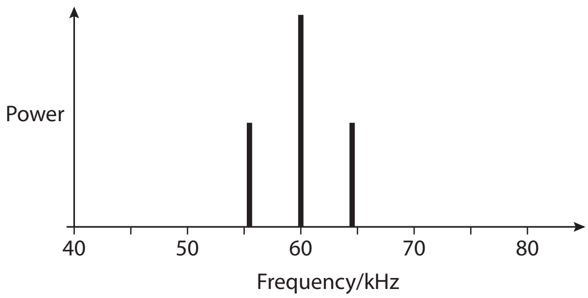
1. **(c)** **(i)** 6 μs **(ii)** 9 μs

2. **(a)** 4 kHz

(b) modulation index = 1

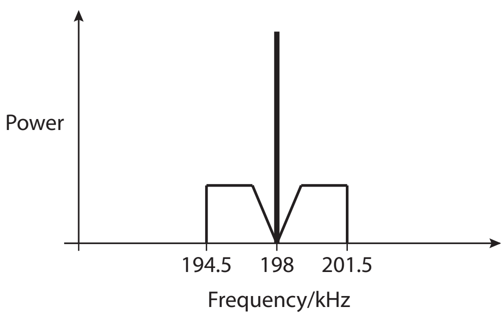
(c) 1 *V*

(d)

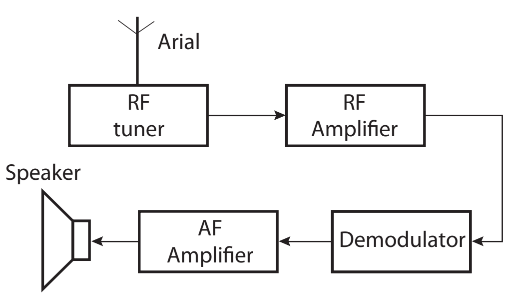


3. **(a)** 7 kHz

(b)



(c)



(d) To increase the amplitude of the signal.

(e) To remove the carrier signal.

4. **(a)** Inverting

(b) 25

5. **(a)**

Time/ms	Voltage level	Binary
0	3	011
2	1	001
4	3	011
6	5	101
8	6	110
10	5	101
12	3	011

(b) Less affected by noise, or can be processed by computer.

6. **(a)** 345789

(b) Because a phone in one cell can receive signals from all the adjoining cells.

(c) As the signal from cell 5 gets stronger the call will be transferred from the transmitter in cell 6 to the one in cell 5.

(d) They don't last long so must be disposed of.

Chapter 12

1. 9.7×10^{14} Hz

2. **(a)** 1.9×10^{15} Hz **(b)** 2.42×10^{14} Hz

3. 4.13 keV

4. visible 6.98×10^{14} Hz

5. radio 8.00×10^{7} Hz

6. infrared 3.00×10^{13} Hz

7. X-ray 3.00×10^{17} Hz

8. 3.98×10^{-2} W m^{-2}

9. 3.75×10^{26} W

10. 41.4 kV

11. 17.8 keV

12. 2.1×10^{-11} m

13. **(a)** strong **(b)** weak

(c) strong

14. **(a)** destructive **(b)** constructive

(c) constructive

15. **(a)** 1 cm **(b)** 1.5 cm **(c)** 0.5 cm

16. 450 cm

17. 562.5 cm

18. 9 cm

19. 6 cm

20. 150 nm

21. 95 nm

22. **(a)** no change **(b)** 387 nm **(c)** 97 nm

23. 104 nm

24. **(a)** 7.4×10^{-4} rad **(b)** 3.7×10^{-5} m

(c) 2.0×10^{-4} m **(d)** 3.1×10^{-4} m

25. **(a)** 3.3 μm **(b)** 12.2°

26. 1.44×10^{-10} m

27. 4.16×10^{-12} m

28. 2.03×10^{-10} m

29. **(a)** 15 cm **(b)** 6.67 dioptres

30. 25 cm

31. **(a)** 15 cm **(b)** real **(c)** 0.5

32. 6.67 cm

33. **(a)** $^{-}7.5$ cm **(b)** virtual

(c) 1.5

34. **(a)** 5 m **(b)** 5.05 cm

(c) 0.01 **(d)** 0.01m

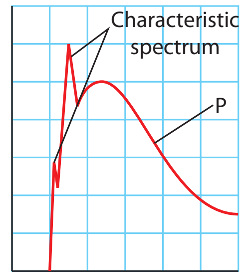
35. **(a)** 6.67 cm **(b)** 0.33

36. 8.75×10^{-3} rads

37. 4×10^{-3} rads
38. 4.16 cm
39. 6
40. (a) 3 cm (b) 4.17 cm (c) 7.17 cm
41. (a) 10 (b) 110 cm
42. 5 cm

Practice questions

1. (b) (i) $v = -25$ cm, so distance is 25 cm
(ii) 4.0 cm
2. (c) at infinity
(e) $v = 2.04$ cm; beyond eyepiece lens/between eyepiece lens and eye
4. (d) $m = \frac{-30}{75} = -0.4$
5. (c) 1.58 mm
6. (a) interference; 1 max
Award no marks for "diffraction".
(b) (i) n is the refractive index of the oil; the wavelength in the medium is reduced / depends on n; 2 max
(ii) $t = \frac{650}{(4 \times 1.45)} = 110$ nm; 1 max
*Award [**0**] for use of incorrect refractive index.*
(iii) *Weak answer giving some indication.*
that the rays reflected off the bottom of the oil film now shift / change in phase by $\frac{\lambda}{2}(\pi)$
A better defined statement.
that initially the rays reflected off the bottom of the oil film (on water), did not suffer a phase change;
but now do shift in phase by $\frac{\lambda}{2}(\pi)$ 2 max
[6]
7. (a) surfaces are flat (to less than $\frac{\lambda}{4}$); 1
Expect some reference to the wavelength of the light 2 not necessarily $\frac{1}{4}\lambda$.
(b) no geometrical path difference;
but π phase change on reflection from surface of less (optically) dense medium;
so destructive interference; 3
(c) fringe separation corresponds to change in thickness of $\frac{1}{2}\lambda$;
there are $\frac{90}{1.4} = 64$ fringes;
thickness $64 \times 5.89 \times 10^{-7} \times \frac{1}{2}$
$= 1.8(8) \times 10^{-5}$ m; 3
*Award [**2 max**] if fringe separation corresponds to λ, not $\frac{1}{2}\lambda$, answer is 3.8×10^{-5} m.*
(d) (i) there are many different wavelengths / *OWTTE*; 1
(ii) complete destructive interference for one wavelength / colour;
remaining wavelengths give coloured appearance; 2
[10]
8. (a)



3 max
*Award [**1**] for each label up to [**3 max**].*
Only one characteristic spectrum peak needs to be labelled.
(b) $f = \frac{Ve}{h}$;
$= \frac{25 \times 10^3 \times 1.6 \times 10^{-19}}{6.63 \times 10^{-34}}$
$= 6.0 \times 10^{18}$ Hz; 2 max
[5]

Chapter 13

1. (a) $8.5\ \text{m s}^{-1}$ (b) 10 m (c) 170m
2. (a) (i) 11.5 s (ii) 27.5 s
(b) (i) 11.5 s (ii) 27.5 s
(c) different frames
3. (a) 1.4×10^{-5} s
(b) 25
4. (a) 2.6 m (b) 3.48 m (c) 0.66 c
5. 6.71×10^9 m
6. 0.866 c
7. 0.99 c
8. 0.85 c
9. 0.96 c
10. 0.045 c
11. 1.17 MeV
12. 2 MV
13. 375 MV
14. 134 MeVc^{-2}
15. (a) 1183 MeVc^{-1} (b) 0.986 c
16. (a) 100.5 MeV (b) 250.5 MeV
(c) 200.5 MeVc^{-1}
17. (a) 950 MeV (b) 12 MeV
(c) 12 MV (d) 0.16 c
18. 8.56×10^3 Hz
19. 29.6 km
20. (a) 60.0008 s (b) 71.5 s
(c) 84.3 s

Practice questions

1. (a) *proper time:*
the time interval measured by an observer of an event that happens at the same place according to that observer;
proper length:
the length of an object as measured by an observer who is at rest relative to the object; 2
Do not look for precise wording but look for the understanding of the quantities in the sense of the words.

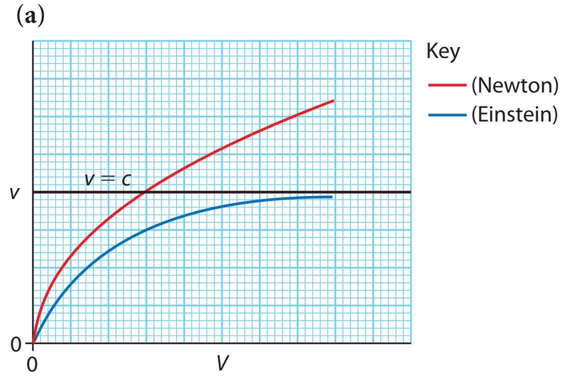
(b) (i) no they will not appear to be simultaneous;
Look for a discussion along the following lines.
Carmen sees Miguel move away from the signal from A and since Miguel receives the two signals at the same time;
and since the speed of light is independent of the motion of the source;
Carmen will see the light from A first / light from B will reach Carmen after light from A / *OWTTE*; 4 max

(ii) $\gamma = 2$;
to give $u = 0.87c$ (2.6×10^8 m s^{-1}); 2

(iii) both measure the correct distance;
SR states that there is no preferred reference system / laws of physics are the same for all inertial observers;
OWTTE; 2 max

[10]

2. (a)



correct general shape;
asymptotic to *c*; 2 max

(b) as the speed of the electrons increases SR predicts that the mass of the electrons will increase;
SR also predicts that at speed *c* the mass will be infinite;
so effectively the electrons can never reach the speed of light; 3

Look for an answer that shows that mass increases and why the electrons cannot travel at the speed of light. They might quote $m = \gamma m_0$ and this is fine.

(c) (i) $\gamma = \dfrac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}}$
to give $\gamma = 4.1$;
$m = \gamma m_0 = 4.1 \times 0.51 = 2.1$ MeVc^{-2};
Accept $m = 3.7 \times 10^{-30}$ *kg.*
could also solve from KE = 1.5 MeV;
rest mass 0.51 = MeVc^{-2};
therefore total mass = 2.1 MeVc^{-2} 3 max

(ii) $E = mc^2$;
= 2.1 MeV; 2
Accept 3.20×10^{-13} *J.* [10]

3. (a) frame moving with constant velocity / frame in which Newton's
first law is valid; 1

(b) $T_0 = \dfrac{2D}{c}$; 1

(c) (i) light reflected off mirror when midway between F and R; 1

(ii) FR = vT; 1

(iii) $(\frac{1}{2}L)^2 = D^2 + (\frac{1}{2}vT)^2$;
$L = \sqrt[2]{\{D^2 + (\frac{1}{2}vT)^2\}}$; 2

(iv) $T = = \dfrac{\sqrt[2]{\{D^2 + (\frac{1}{2}vT)^2\}}}{c}$;
$c^2T^2 = 4\{D^2 + (\frac{1}{2}vT)^2\}$;
use of $4D^2 = c^2T_0^{\,2}$;
hence $T = T_0\sqrt{\left(1 - \dfrac{v^2}{c^2}\right)}$; 4

[10]

4. (a) rest mass energy is the energy that is needed to create the particle at rest / reference to $E_0 = m_0c^2$;
total energy is the addition of the rest energy and everything else (kinetic *etc.*) / reference to mass being greater when in motion / $E = mc^2$;
2 max

(b) realization that betas are electrons;
so $me = 0.511$ MeVc^{-2};
$\lambda = \dfrac{2.51}{0.511}$; (= 4.91) 3
Ignore any spurious calculation from Lorentz factor equation here as the use of this equation is rewarded below.

(c) (i) correct substitution into Lorentz factor equation;
to give $v = 0.979c = 2.94 \times 10^8$ m s^{-1}; 2

(ii) correct substitution into speed $= \dfrac{\text{distance}}{\text{time}}$;
to give time = 1.26 ns; 2

(d) **(i)** the detector / the laboratory / *OWTTE*; 1
(ii) same answer as **(c) (i)** $= 2.94 \times 10^8\ \mathrm{ms}^{-1}$; 1
(iii) realization that length contraction applies;
distance $= \frac{37}{\gamma} = 7.5$ cm; 2
[13]

5. **(a)** the speed of light <u>in vacuum</u> is the same for all inertial observers;
the laws of Physics are the same in all <u>inertial</u> frames of reference; 2
(b) **(i)** this faster than light speed is not the speed of any physical object / inertial observer and so is not in violation of the theory of SR; 1
(ii) $u' = \dfrac{u - v}{1 - \dfrac{uv}{c^2}}$ with $v = -0.80\,c$
and $u = 0.80\,c$ so that
$u' = \dfrac{0.80\,c + 0.80\,c}{1 + \dfrac{0.80\,c \times 0.80\,c}{c^2}}$;
$u = \dfrac{1.60\,c}{1.64}$;
$u = 0.98\,c$; 3
[6]

6. **(a)** *Award [2] for good understanding and [1* ***max****] for some understanding.*
a means by which the position of an object can be located / *OWTTE*;
some detail *e.g.* reference to origin/axes; 2
Answers will be open-ended.
(b) $c - v$; 1
(c) c; 1 max
(d) $u' = \dfrac{u - v}{1 - \dfrac{cv}{c^2}}$;
substitute $u = c$ to get $u' = \dfrac{c - v}{1 - \dfrac{cv}{c^2}}$;
$= \dfrac{c - v}{1 - \dfrac{v}{c}}; = \dfrac{c(c - v)}{c - v} = c$; 3
Accept answers using + instead of −.
Award [1] for recognition of correct formula to use and [1] for correct substitution and [1] for at least some arithmetic.
(e) **(i)** time interval of an event that is observed to happen at the same place / *OWTTE*; 1
(ii) $\gamma = 2.0$;
$2.0 = \dfrac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}}$;
arithmetic to give $v = 0.87\,c$; 3
[11]

Chapter 14

1. 8×10^{-6} N
2. 1.5
3. 1.2×10^{-5} N
4. 4 Pa
5. 20
6. **(a)** $7.96 \times 10^{-5}\ \mathrm{W\,m^{-2}}$
(b) 79 dB
7. **(a)** 10^{-2} W **(b)** 0.25 W **(c)** 114 dB
8. 78 500
9. **(a)** 10^{-2} W **(b)** 0.13 W
10. **(a)** $5.6 \times 10^{-2}\ \mathrm{mm^{-1}}$ **(b)** 12.4 mm
11. **(a)** $0.693\ \mathrm{mm^{-1}}$ **(b)** $0.0625\ \mathrm{kW\,m^{-2}}$
12. **(a)** $0.153\ \mathrm{mm^{-1}}$ **(b)** 4.5 mm
13. **(a)** $1.63 \times 10^6\ \mathrm{kg\,m^2\,s^{-1}}$
(b) $7.18 \times 10^6\ \mathrm{kg\,m^2\,s^{-1}}$
(c) $1.33 \times 10^6\ \mathrm{kg\,m^2\,s^{-1}}$
14. Fat and Bone
15. **(a)** 4.5 cm **(b)** 4.5 cm
16. **(a)** 2.88×10^{-7} J **(b)** 2.88×10^{-8} J
(c) 4.1×10^{-10} Gy
17. 0.14 J
18. 7 mJ
19. **(a)** 70 J **(b)** 4.38×10^{14}
(c) 4.38×10^{15} **(d)** 1.22×10^{12} Bq
20. 0.2 days
21. 5 ltr

Practice questions

1. **(a)** 15-30 Hz to 15-20 kHz; 1
(b) sounds of different frequency force different hair cells to vibrate;
at different amplitude depending on the length / thickness / stiffness of the hair cells;
the (electrical) signals sent to the brain from the different vibrating receptors allow the brain to distinguish different frequencies; 3
As candidates are unlikely to answer at this level of detail, be generous and award marks accordingly if they show ***understanding*** *of each of the processes involved.*
(c) **(i)** $\beta = 10 \log \dfrac{I}{10^{-12}}$
$\beta = 10 \log \dfrac{2.7 \times 10^{-5}}{10^{-12}}$;
$\beta = 74$ dB; 2
(ii) the response of the ear is logarithmic;
the sound intensity level β is defined in terms of the logarithm of sound intensity;

so equal changes in β correspond to equal changes in *ratios* of intensity; 3

[9]

2. **(a)** *conductive*: vibrations/sound does not reach the inner ear;
sensory: the inner ear does not pass impulses to the brain; 2

(b) intensity level in decibels $= 10 \log_{10}\frac{I}{I_0}$ and $I_0 = 10^{-12}\ \text{W m}^{-2}$;
where I is the measured intensity; 2
Allow **[1]** *for "sound intensity in W m^{-2} is related to sound intensity level by a logarithmic scale" / OWTTE.*

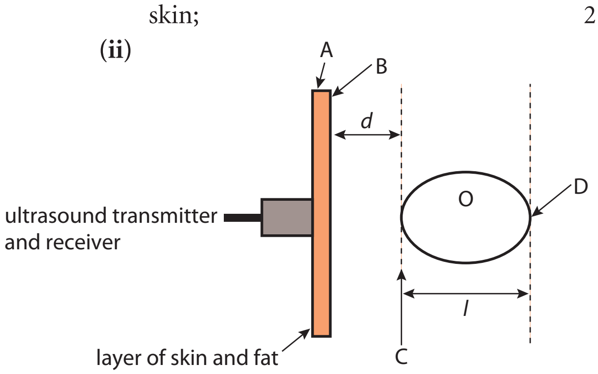
(c) *Frederick*:
conductive – the uniform loss with frequency suggests damage to the ear;
damage could be caused by ear infection, perforation of eardrum *etc.*;
Susanna:
sensory – the hearing loss is increasing with increasing frequency;
damage could be due to old age / continual exposure to excessive noise / disease; 4
Also accept could be conductive loss. Award **[0]** *for just stating the correct loss.*

[8]

3. **(a)** 1 MHz → 20 MHz; 1

(b) **(i)** to ensure that no air is trapped between transmitter and skin;
otherwise nearly all the transmitted pulse will be reflected at the surface of the skin; 2

(ii)



A and B correct;
C and D correct; 2

(iii) pulse takes 50 μs to travel $2d$;
therefore $d = \left(\frac{1.5 \times 50}{2}\right) \times 10^{-3}$ m;
to give $d = 38$ mm;
similarly $l = \left(\frac{1.5 \times 175}{2}\right) \times 10^{-3}$ m
$= 130$ mm; 4
Allow for ECF here e.g. if "d" is marked as being between A and B.

(c) B-scan gives a three-dimensional image;
OWTTE; 1 max

(d) *advantage*:
non-ionising *(not as harmful as X-rays / OWTTE)*;
Any ***one*** *of the following:*
disadvantages:
small depth of penetration;
limit to size of objects that can be imaged;
blurring of images due to reflection at boundaries; 2 max

[12]

4. **(a)** **(i)** X-rays;
because they can easily distinguish between flesh and bone to get a clear image of the fracture; 2

(ii) ultrasound;
because it gives reasonably clear images in the womb without harmful radiation; 2

(b) **(i)** the half-value thickness is that thickness of lead which (for this particular beam);
will reduce the intensity of the (transmitted) beam by 50 %; 2

(ii) the half-value thickness corresponds to an intensity of 10 units;
and so equals 4 mm; 2

(iii) the transmitted intensity must be
$20\ \% \times 20 = 4$ units;
corresponding to a thickness of lead of about 9.3 ($\pm$0.2) mm; 2

(iv) the transmitted intensity must be
$(1 - 0.8) \times 20 = 4$ units;
using $4 = 20(0.5)^{x/8} \Rightarrow (0.5)^{x/8} = 0.20$;
we find a thickness of 18.6 ($\pm$1) mm; 3
or
the transmitted intensity must be
$(1 - 0.8) \times 20 = 4$ units;
by drawing a second graph corresponding to the half–value thickness of 8 mm;
and finding the thickness corresponding to a transmitted intensity of 4 units of about 18.6 ($\pm$1) mm; 3

[13]

5. **(a)** a general statement is adequate here, *e.g.* it allows doses of different types of radiation to be compared for their biological effects;
of course more detailed statements are acceptable, *e.g.* quality factor/RBE compares the biological effectiveness of a given type of radiation to that of [200 keV] X-rays; 1 max

(b) absorbed dose $= \dfrac{\text{dose equivalent}}{\text{quality factor}}$

$= \dfrac{240\ \text{J kg}^{-1}}{14} = 17.1\ \text{J kg}^{-1}$

total energy absorbed

$= 17 \times 0.015 = 0.25$ J;

= number per second × energy per proton × time;

exposure time

$= \dfrac{0.257}{1.8 \times 10^{10} \times 4.0 \times 10^{6} \times 1.6 \times 10^{-19}}$

$= 22(.3)$ s;

Watch for ECF. 4 max

[5]

6. (a) biological half−life is the time it takes the body to eject by natural bodily processes; half of an ingested sample of a radioactive isotope; 2 max

To award [2] some mention must be made of the general or specific method by which the amount of the isotope in the body is reduced.

(b) $\dfrac{1}{T_E} = \dfrac{1}{21} + \dfrac{1}{8}$;

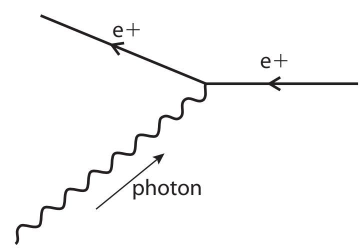
to give $TE = 5.8$ days; 2

(c) because of its short physical half life it is much less likely to cause damage to the thyroid gland / because person is radioactive for a shorter time / because total dose received would be smaller / *OWTTE*; 1

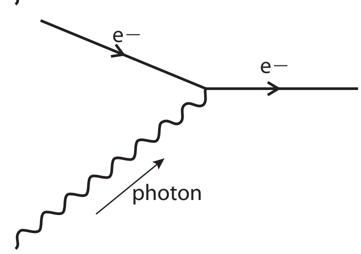
[5]

Chapter 15

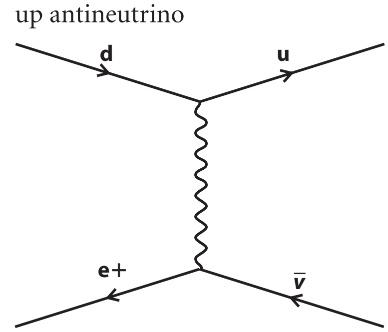
1. 5.3×10^{-20} N s
2. 938 MeV
3. Lepton
   Meson
   Baryon
   Lepton
4. yes
5. no
6. no
7. yes
8. yes
9. no
10. no
11. 1.876 GeV
12. (a) 1000 eV (b) 1.6×10^{-16} J
    (c) 1.7×10^{-23} N s (d) 3.9×10^{-11} m
13. (a) 1.07×10^{-17} N s (b) 6.22×10^{-17} m
14. 3.86 GeV
15. (a) 2.01 GeV (b) 1.22 GeV (c) 0.28 GeV
16. (a) 3.752 GeV (b) 6.93 GeV
    (c) 5.63 GeV (d) 0.938 GeV
17. (a) 3.36×10^{8} J (b) 13 kmhr^{-1}
18. 1.12×10^{-6} J
19. 1.12×10^{6} J
20. 1500 ms^{-1}
21. 1.12×10^{15} W
22. B – positron
    A – electron
23. (a) weak (b) not weak (c) weak
24. (a) yes (b) yes (c) no
25. (a) $d\bar{u}$ (b) sss (c) ssd (d) ssu
26. $d \rightarrow u$
27. no
28. yes
29. yes
30. no
31. gluon
32. W
33. (a)



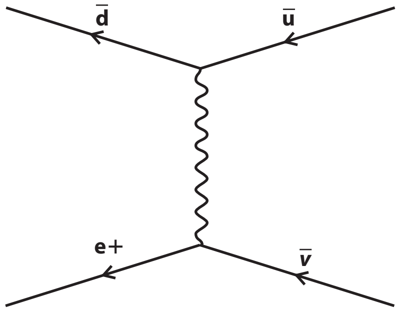
(b)



34. (a) down quark
    (b) weak
    (c) up quark
    (d) up quark, antineutrino, electron
    (e) beta decay
35. (a) up antineutrino



(b) anti down, antineutrino



36. 8.3×10^{15} K

37. 2.07×10^{-8} J; 1.3×10^{11} MeV

38. Before 3 min, KE of particles enough to break up He.

Practice questions

1. **(i)** muon lepton number / electron lepton number; 1

(ii) baryon number; 1

(iii) baryon number / electric charge; 1

(b) there are eight gluons involved in the strong interaction; 1

Accept just the name gluons or just mesons.

[4]

2. **(a)** Hadron; (*Award this mark for "bald" statement and if reason is wrong.*)

any sensible justification; 2

e.g. "contains two quarks" or "hadrons are either Baryons or mesons".

(b) any combination of three quarks;

correct answer: UUD; 2

(c) attempt (even if unsuccessful) to balance quarks left and right;

to get: $\begin{pmatrix} s \\ u \end{pmatrix} + \begin{pmatrix} u \\ u \\ d \end{pmatrix} \rightarrow \begin{pmatrix} d \\ s \end{pmatrix} + \begin{pmatrix} u \\ s \end{pmatrix} + \begin{pmatrix} s \\ s \\ s \end{pmatrix}$

correct discussion on how the equation balances for all quark types; 2 max

e.g. compare numbers of quarks on LHS and RHS:

$u: -1\ (1 + 1) \rightarrow 1$

$d: 1 \rightarrow 1$

$s: 1 \rightarrow -1 - 1 + (1 + 1 + 1)$

[6]

3. **(a)** **(i)** colour force / weak force; 1

(ii) gluon / charged vector boson / W boson; 1

(b) in the interaction $\bar{\nu} + p = n + e^+$ charge, lepton number and baryon number are conserved / all conservation laws are conserved;

in the interaction $\nu + p = n + e^+$ charge and baryon number are conserved / all conservation laws except lepton number are conserved;

lepton number, +1 on the left −1 on the right; 3 max

*Essentially look for some detail of the conservation laws and some substantiation of the violation of lepton number to achieve **[3 max]**.*

[5]

4. **(a)** Short lived, allowed to exist due to uncertainty principle.

(b) One particle emits a photon causing it to recoil in the opposite direction. Absorbed by second particle causing force. Photon has zero rest mass so force is infinite.

(c) Gravity _____ Infinite

_____ _____ Infinite

Weak _____ Short

Strong Gluon Short

5. **(a)** U, U and S;

(b) Colour force/strong interaction;

due to interchange of gluons;

(c) Charm, top (truth) and bottom (beauty).

6. **(i)** Lepton number not conserved $n \rightarrow p^+ + e^- + \bar{\nu}$

(ii) Baryon number not conserved $\Lambda^\circ \rightarrow p^+ + \pi^-$

(iii) Momentum not conserved $e^- + e^- \rightarrow \gamma + \gamma$

7. **(a)** $\pi^+ + \mathrm{p} \rightarrow \Lambda^\circ + K^\circ + \pi^+ + \pi^+$

(b) Charge $1 + 1 = 0 + 1 + 1$

Baryon number $0 + 1 = 1 + 0 + 0 + 0$

Strangeness $0 + 0 = -1 + 1 + 0 + 0$

(c) Initial K.E. converted in the collision.

(d) Any one of mass / energy, momentum and angular momentum must also be conserved.

Index

Page numbers in *italics* refer to information boxes and diagrams.